(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 April 2002 (18.04.2002)

PCT

(10) International Publication Number
WO 02/30951 A2

(51) International Patent Classification[7]: C07K

(21) International Application Number: PCT/US01/42726

(22) International Filing Date: 15 October 2001 (15.10.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/240,211 13 October 2000 (13.10.2000) US

(71) Applicant *(for all designated States except US)*: **GENAISSANCE PHARMACEUTICALS, INC.** [US/US]; Five Science Park, New Haven, CT 06511 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **HAN, Jin-Hua** [CN/US]; 569B Prospect Street, New Haven, CT 06511 (US). **KLIEM, Stefanie, E.** [DE/DE]; Kiefernweg 37, 61440 Oberusel (DE). **SANCHIS, Angela** [US/US]; 7 Trumbull Street, Apt. 4, New Haven, CT 06511 (US).

(74) Agents: **FIELD, Gisela, M.** et al.; Genaissance Pharmaceuticals, Inc., Five Science Park, New Haven, CT 06511 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/30951 A2

(54) Title: HAPLOTYPES OF THE CYP1B1 GENE

(57) Abstract: Novel genetic variants of the Cytochrome P450, Subfamily I (Dioxin-Inducible), Polypeptide 1 (Glaucoma 3, Primary Infantile) (CYP1B1) gene are described. Various genotypes, haplotypes, and haplotype pairs that exist in the general United States population are disclosed for the CYP1B1 gene. Compositions and methods for haplotyping and/or genotyping the CYP1B1 gene in an individual are also disclosed. Polynucleotides defined by the haplotypes disclosed herein are also described.

HAPLOTYPES OF THE CYP1B1 GENE

RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application Serial No. 60/240,211 filed October 13, 2000.

FIELD OF THE INVENTION

This invention relates to variation in genes that encode pharmaceutically-important proteins. In particular, this invention provides genetic variants of the human cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene and methods for identifying which variant(s) of this gene is/are possessed by an individual.

BACKGROUND OF THE INVENTION

Current methods for identifying pharmaceuticals to treat disease often start by identifying, cloning, and expressing an important target protein related to the disease. A determination of whether an agonist or antagonist is needed to produce an effect that may benefit a patient with the disease is then made. Then, vast numbers of compounds are screened against the target protein to find new potential drugs. The desired outcome of this process is a lead compound that is specific for the target, thereby reducing the incidence of the undesired side effects usually caused by activity at non-intended targets. The lead compound identified in this screening process then undergoes further *in vitro* and *in vivo* testing to determine its absorption, disposition, metabolism and toxicological profiles. Typically, this testing involves use of cell lines and animal models with limited, if any, genetic diversity.

What this approach fails to consider, however, is that natural genetic variability exists between individuals in any and every population with respect to pharmaceutically-important proteins, including the protein targets of candidate drugs, the enzymes that metabolize these drugs and the proteins whose activity is modulated by such drug targets. Subtle alteration(s) in the primary nucleotide sequence of a gene encoding a pharmaceutically-important protein may be manifested as significant variation in expression, structure and/or function of the protein. Such alterations may explain the relatively high degree of uncertainty inherent in the treatment of individuals with a drug whose design is based upon a single representative example of the target or enzyme(s) involved in metabolizing the drug. For example, it is well-established that some drugs frequently have lower efficacy in some individuals than others, which means such individuals and their physicians must weigh the possible benefit of a larger dosage against a greater risk of side effects. Also, there is significant variation in how well people metabolize drugs and other exogenous chemicals, resulting in substantial interindividual variation in the toxicity and/or efficacy of such exogenous substances (Evans et al., 1999, *Science* 286:487-491). This variability in efficacy or toxicity of a drug in genetically-diverse patients makes many drugs ineffective or even dangerous in certain groups of the population, leading to the failure of

such drugs in clinical trials or their early withdrawal from the market even though they could be highly beneficial for other groups in the population. This problem significantly increases the time and cost of drug discovery and development, which is a matter of great public concern.

It is well-recognized by pharmaceutical scientists that considering the impact of the genetic variability of pharmaceutically-important proteins in the early phases of drug discovery and development is likely to reduce the failure rate of candidate and approved drugs (Marshall A 1997 *Nature Biotech* **15**:1249-52; Kleyn PW et al. 1998 *Science* **281**: 1820-21; Kola I 1999 *Curr Opin Biotech* **10**:589-92; Hill AVS et al. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 62-76; Meyer U.A. 1999 in *Evolution in Health and Disease* Stearns SS (Ed.) Oxford University Press, New York, pp 41-49; Kalow W et al. 1999 *Clin. Pharm. Therap.* **66**:445-7; Marshall, E 1999 *Science* **284**:406-7; Judson R et al. 2000 *Pharmacogenomics* **1**:1-12; Roses AD 2000 *Nature* **405**:857-65). However, in practice this has been difficult to do, in large part because of the time and cost required for discovering the amount of genetic variation that exists in the population (Chakravarti A 1998 *Nature Genet* **19**:216-7; Wang DG et al 1998 *Science* **280**:1077-82; Chakravarti A 1999 *Nat Genet* **21**:56-60 (suppl); Stephens JC 1999 *Mol. Diagnosis* **4**:309-317; Kwok PY and Gu S 1999 *Mol. Med. Today* **5**:538-43; Davidson S 2000 *Nature Biotech* **18**:1134-5).

The standard for measuring genetic variation among individuals is the haplotype, which is the ordered combination of polymorphisms in the sequence of each form of a gene that exists in the population. Because haplotypes represent the variation across each form of a gene, they provide a more accurate and reliable measurement of genetic variation than individual polymorphisms. For example, while specific variations in gene sequences have been associated with a particular phenotype such as disease susceptibility (Roses AD *supra*; Ulbrecht M et al. 2000 *Am J Respir Crit Care Med* **161**: 469-74) and drug response (Wolfe CR et al. 2000 *BMJ* **320**:987-90; Dahl BS 1997 *Acta Psychiatr Scand* **96** (Suppl 391): 14-21), in many other cases an individual polymorphism may be found in a variety of genomic backgrounds, i.e., different haplotypes, and therefore shows no definitive coupling between the polymorphism and the causative site for the phenotype (Clark AG et al. 1998 *Am J Hum Genet* **63**:595-612; Ulbrecht M et al. 2000 *supra*; Drysdale et al. 2000 *PNAS* **97**:10483-10488). Thus, there is an unmet need in the pharmaceutical industry for information on what haplotypes exist in the population for pharmaceutically-important genes. Such haplotype information would be useful in improving the efficiency and output of several steps in the drug discovery and development process, including target validation, identifying lead compounds, and early phase clinical trials (Marshall et al., *supra*).

One pharmaceutically-important gene for the treatment of breast cancer and primary congenital glaucoma is the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene or its encoded product. CYP1B1 belongs to the multigene cytochrome P450 superfamily, a group of monomeric heme-thiolate monooxygenases that participate in an electron transport pathway as part of phase 1 cellular metabolism. These enzymes are induced

by polycyclic aromatic hydrocarbons (PAH) and 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), which are widespread chemical pollutants known to be potent carcinogen and tumor-promoting agents in rodents (Sutter et al. *J. Biol.Chem.* 1994. 269: 13092-13099). CYP1B1 oxidizes a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics.

Specifically, CYP1B1 is a key enzyme involved in the production of potentially carcinogenic estrogen metabolites and the activation of environmental carcinogens. CYP1B1 is the predominant member of the CYP1 family expressed in normal breast tissue and breast cancer, and studies indicate that genetic differences in CYP1B1 could account for interindividual differences in steroid receptor expression that may be functionally important in breast cancer pathogenesis (Bailey LR, et al. *Cancer Res* 1998 Nov 15;58(22):5038-41). Additionally, linkage studies of candidate genes identified in the critical region of 2p21, where a major gene for primary congenital glaucoma (PCG) had been mapped, CYP1B1 was discovered as the first example of the cytochrome P450 superfamily in which mutations resulted in a primary developmental defect (Stoilov et al. *Hum. Molec. Genet.* 1997. 6: 641-647). Biochemical studies have suggested that CYP1B1 participates in the metabolism of an as-yet-unknown biologically active molecule that is a participant in eye development, and that protein variants result in clouding of the cornea and inhibition of regulation of aqueous humor secretion, the two major diagnostic criteria for PCG (Schwartzman et al. *Proc. Nat. Acad. Sci.* 1987. 84: 8125-8129). Mutations in CYP1B1 are implicated therefore as direct causative factors in PCG, a recessive disorder characterized by large ocular globes resulting from increased intraocular pressure.

The cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) gene is located on chromosome 2p21 and contains 3 exons that encode a 543 amino acid protein. A reference sequence for the CYP1B1 gene is shown in the contiguous lines of Figure 1 (Genaissance Reference No. 1834513; SEQ ID NO: 1). Reference sequences for the coding sequence (GenBank Accession No. NM_000104.1) and protein are shown in Figures 2 (SEQ ID NO: 2) and 3 (SEQ ID NO: 3), respectively.

Six polymorphisms of the CYP1B1 gene have been previously identified. These single nucleotide polymorphisms correspond to the sites named PS8, PS9, PS15, PS17, PS18 and PS20 herein. Specifically, the variation which corresponds to PS8 consists of a cytosine or guanine at nucleotide position 2610 in Figure 1 (NCBI SNP ID: rs10012). This polymorphism is expressed in the coding sequence at nucleotide position 142 in Figure 2, and results in an amino acid variation of arginine or glycine at position 48 in Figure 3. The variation which corresponds to PS9 consists of a guanine or thymine at nucleotide position 2823 in Figure 1 (NCBI SNP ID: rs1056827). This polymorphism is expressed in the coding sequence at nucleotide position 355 in Figure 2, and results in an amino acid variation of alanine or serine at position 119 in Figure 3. The variation which corresponds to PS15 consists of a guanine or cytosine at nucleotide position 6798 in Figure 1 (NCBI SNP ID: rs1056836). This polymorphism is expressed in the coding sequence at nucleotide position 1294 in Figure 2, and results in an amino acid variation of valine or leucine at position 432 in Figure

3. The variation which corresponds to PS17 consists of a thymine or cytosine at nucleotide position 6851 in Figure 1 (NCBI SNP ID: rs1056837). This polymorphism is expressed in the coding sequence at nucleotide position 1347 in Figure 2, but has no effect on the amino acid sequence. The variation which corresponds to PS18 consists of a adenine or guanine at nucleotide position 6862 in Figure 1 (NCBI SNP ID: rs1800440). This polymorphism is expressed in the coding sequence at nucleotide position 1358 in Figure 2, and results in an amino acid variation of asparagine or serine at position 453 in Figure 3. Finally, the variation which corresponds to PS20 consists of a cytosine or guanine at nucleotide position 7254 in Figure 1 (NCBI SNP ID: rs1799885).

Because of the potential for variation in the CYP1B1 gene to affect the expression and function of the encoded protein, it would be useful to know whether additional polymorphisms exist in the CYP1B1 gene, as well as how such polymorphisms are combined in different copies of the gene. Such information could be applied for studying the biological function of CYP1B1 as well as in identifying drugs targeting this protein for the treatment of disorders related to its abnormal expression or function.

SUMMARY OF THE INVENTION

Accordingly, the inventors herein have discovered 14 novel polymorphic sites in the CYP1B1 gene. These polymorphic sites (PS) correspond to the following nucleotide positions in Figure 1: 1000 (PS1), 1071 (PS2), 1279 (PS3), 1294 (PS4), 1405 (PS5), 2391 (PS6), 2393 (PS7), 2969 (PS10), 3134 (PS11), 3488 (PS12), 6488 (PS13), 6602 (PS14), 6769 (PS16) and 7179 (PS19). The polymorphisms at these sites are cytosine or thymine at PS1, thymine or cytosine at PS2, guanine or adenine at PS3, thymine or cytosine at PS4, cytosine or thymine at PS5, cytosine or thymine at PS6, cytosine or thymine at PS7, cytosine or adenine at PS10, guanine or cytosine at PS11, cytosine or thymine at PS12, thymine or cytosine at PS13, adenine or guanine at PS14, cytosine or guanine at PS16 and guanine or adenine at PS19. In addition, the inventors have determined the identity of the alleles at these sites, as well as at the previously identified sites at nucleotide positions 2547 (PS8), 2760 (PS9), 6735 (PS15), 6788 (PS17), 6799 (PS18) and 7191 (PS20), in a human reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: African descent, Asian, Caucasian and Hispanic/Latino. From this information, the inventors deduced a set of haplotypes and haplotype pairs for PS1-PS20 in the CYP1B1 gene, which are shown below in Tables 5 and 4, respectively. Each of these CYP1B1 haplotypes constitutes a code that defines the variant nucleotides that exist in the human population at this set of polymorphic sites in the CYP1B1 gene. Thus each CYP1B1 haplotype also represents a naturally-occurring isoform (also referred to herein as an "isogene") of the CYP1B1 gene. The frequency of each haplotype and haplotype pair within the total reference population and within each of the four major population groups included in the reference population was also determined.

Thus, in one embodiment, the invention provides a method, composition and kit for

genotyping the CYP1B1 gene in an individual. The genotyping method comprises identifying the nucleotide pair that is present at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19 in both copies of the CYP1B1 gene from the individual. A genotyping composition of the invention comprises an oligonucleotide probe or primer which is designed to specifically hybridize to a target region containing, or adjacent to, one of these novel CYP1B1 polymorphic sites. A genotyping kit of the invention comprises a set of oligonucleotides designed to genotype each of these novel CYP1B1 polymorphic sites. In a preferred embodiment, the genotyping kit comprises a set of oligonucleotides designed to genotype each of PS1-PS20. The genotyping method, composition, and kit are useful in determining whether an individual has one of the haplotypes in Table 5 below or has one of the haplotype pairs in Table 4 below.

The invention also provides a method for haplotyping the CYP1B1 gene in an individual. In one embodiment, the haplotyping method comprises determining, for one copy of the CYP1B1 gene, the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19. In another embodiment, the haplotyping method comprises determining whether one copy of the individual's CYP1B1 gene is defined by one of the CYP1B1 haplotypes shown in Table 5, below, or a sub-haplotype thereof. In a preferred embodiment, the haplotyping method comprises determining whether both copies of the individual's CYP1B1 gene are defined by one of the CYP1B1 haplotype pairs shown in Table 4 below, or a sub-haplotype pair thereof. Establishing the CYP1B1 haplotype or haplotype pair of an individual is useful for improving the efficiency and reliability of several steps in the discovery and development of drugs for treating diseases associated with CYP1B1 activity, e.g., breast cancer and primary congenital glaucoma.

For example, the haplotyping method can be used by the pharmaceutical research scientist to validate CYP1B1 as a candidate target for treating a specific condition or disease predicted to be associated with CYP1B1 activity. Determining for a particular population the frequency of one or more of the individual CYP1B1 haplotypes or haplotype pairs described herein will facilitate a decision on whether to pursue CYP1B1 as a target for treating the specific disease of interest. In particular, if variable CYP1B1 activity is associated with the disease, then one or more CYP1B1 haplotypes or haplotype pairs will be found at a higher frequency in disease cohorts than in appropriately genetically matched controls. Conversely, if each of the observed CYP1B1 haplotypes are of similar frequencies in the disease and control groups, then it may be inferred that variable CYP1B1 activity has little, if any, involvement with that disease. In either case, the pharmaceutical research scientist can, without *a priori* knowledge as to the phenotypic effect of any CYP1B1 haplotype or haplotype pair, apply the information derived from detecting CYP1B1 haplotypes in an individual to decide whether modulating CYP1B1 activity would be useful in treating the disease.

The claimed invention is also useful in screening for compounds targeting CYP1B1 to treat a

specific condition or disease predicted to be associated with CYP1B1 activity. For example, detecting which of the CYP1B1 haplotypes or haplotype pairs disclosed herein are present in individual members of a population with the specific disease of interest enables the pharmaceutical scientist to screen for a compound(s) that displays the highest desired agonist or antagonist activity for each of the CYP1B1 isoforms present in the disease population, or for only the most frequent CYP1B1 isoforms present in the disease population. Thus, without requiring any *a priori* knowledge of the phenotypic effect of any particular CYP1B1 haplotype or haplotype pair, the claimed haplotyping method provides the scientist with a tool to identify lead compounds that are more likely to show efficacy in clinical trials.

Haplotyping the CYP1B1 gene in an individual is also useful to control for genetically-based bias in the design of candidate drugs that target or are metabolized by CYP1B1. For example, for a lead compound that is metabolized by CYP1B1, the pharmaceutical scientist of ordinary skill would be concerned that a favorable efficacy and/or side effect profile shown in a Phase II or Phase III trial may not be replicated in the general population if a higher (or lower) percentage of patients in the treatment group, compared to the general population, have a form of the CYP1B1 gene that makes them genetically predisposed to metabolize the drug more efficiently than patients with other forms of the CYP1B1 gene. Similarly, this pharmaceutical scientist would recognize the potential for bias in the results of a Phase II or Phase III clinical trial of a drug targeting CYP1B1 that could be introduced if individuals whose CYP1B1 gene structure makes them genetically predisposed to respond well to the drug are present in a higher (or lower) frequency in the treatment group than in the control group (Bacanu et al., 2000, *Am. J. Hum. Gen.* **66**:1933-44; Pritchard et al., 2000, *Am. J. Hum. Gen.* **67**: 170-81).

The pharmaceutical scientist can immediately reduce this potential for genetically-base bias in the results of clinical trials of drugs metabolized by or targeting CYP1B1 by practicing the claimed invention. In particular, by determining which of the CYP1B1 haplotypes disclosed herein are present in individuals recruited to participate in a clinical trial of a drug metabolized by or targeting CYP1B1, the pharmaceutical scientist can then assign that individual to the treatment or control group as appropriate to ensure that approximately equal frequencies of different CYP1B1 haplotypes (or haplotype pairs) are represented in the two groups and/or the frequencies of different CYP1B1 haplotypes or haplotype pairs are similar to the frequencies in the general population. Thus, by practicing the claimed invention, the pharmaceutical scientist can more confidently rely on the information learned from the trial, without first determining the phenotypic effect of any CYP1B1 haplotype or haplotype pair.

In another embodiment, the invention provides a method for identifying an association between a trait and a CYP1B1 genotype, haplotype, or haplotype pair for one or more of the novel polymorphic sites described herein. The method comprises comparing the frequency of the CYP1B1 genotype, haplotype, or haplotype pair in a population exhibiting the trait with the frequency of the

CYP1B1 genotype or haplotype in a reference population. A higher frequency of the CYP1B1 genotype, haplotype, or haplotype pair in the trait population than in the reference population indicates the trait is associated with the CYP1B1 genotype, haplotype, or haplotype pair. In preferred embodiments, the trait is susceptibility to a disease, severity of a disease, the staging of a disease or response to a drug. In a particularly preferred embodiment, the CYP1B1 haplotype is selected from the haplotypes shown in Table 5, or a sub-haplotype thereof. Such methods have applicability in developing diagnostic tests and therapeutic treatments for breast cancer and primary congenital glaucoma.

In yet another embodiment, the invention provides an isolated polynucleotide comprising a nucleotide sequence which is a polymorphic variant of a reference sequence for the CYP1B1 gene or a fragment thereof. The reference sequence comprises the contiguous sequences shown in Figure 1 and the polymorphic variant comprises at least one polymorphism selected from the group consisting of thymine at PS1, cytosine at PS2, adenine at PS3, cytosine at PS4, thymine at PS5, thymine at PS6, thymine at PS7, adenine at PS10, cytosine at PS11, thymine at PS12, cytosine at PS13, guanine at PS14, guanine at PS16 and adenine at PS19. In a preferred embodiment, the polymorphic variant comprises one or more additional polymorphisms selected from the group consisting of guanine at PS8, thymine at PS9, cytosine at PS15, cytosine at PS17, guanine at PS18 and guanine at PS20.

A particularly preferred polymorphic variant is an isogene of the CYP1B1 gene. A CYP1B1 isogene of the invention comprises cytosine or thymine at PS1, thymine or cytosine at PS2, guanine or adenine at PS3, thymine or cytosine at PS4, cytosine or thymine at PS5, cytosine or thymine at PS6, cytosine or thymine at PS7, cytosine or guanine at PS8, guanine or thymine at PS9, cytosine or adenine at PS10, guanine or cytosine at PS11, cytosine or thymine at PS12, thymine or cytosine at PS13, adenine or guanine at PS14, guanine or cytosine at PS15, cytosine or guanine at PS16, thymine or cytosine at PS17, adenine or guanine at PS18, guanine or adenine at PS19 and cytosine or guanine at PS20. The invention also provides a collection of CYP1B1 isogenes, referred to herein as a CYP1B1 genome anthology.

In another embodiment, the invention provides a polynucleotide comprising a polymorphic variant of a reference sequence for a CYP1B1 cDNA or a fragment thereof. The reference sequence comprises SEQ ID NO:2 (Fig.2) and the polymorphic cDNA comprises at least one polymorphism selected from the group consisting of adenine at a position corresponding to nucleotide 564, cytosine at a position corresponding to nucleotide 729, cytosine at a position corresponding to nucleotide 1047, guanine at a position corresponding to nucleotide 1161 and guanine at a position corresponding to nucleotide 1328. In a preferred embodiment, the polymorphic variant comprises one or more additional polymorphisms selected from the group consisting of guanine at a position corresponding to nucleotide 142, thymine at a position corresponding to nucleotide 355, cytosine at a position corresponding to nucleotide 1294, cytosine at a position corresponding to nucleotide 1347 and guanine at a position corresponding to nucleotide 1358. A particularly preferred polymorphic cDNA variant

comprises the coding sequence of a CYP1B1 isogene defined by haplotypes 1, 3-11 and 13-20.

Polynucleotides complementary to these CYP1B1 genomic and cDNA variants are also provided by the invention. It is believed that polymorphic variants of the CYP1B1 gene will be useful in studying the expression and function of CYP1B1, and in expressing CYP1B1 protein for use in screening for candidate drugs to treat diseases related to CYP1B1 activity.

In other embodiments, the invention provides a recombinant expression vector comprising one of the polymorphic genomic and cDNA variants operably linked to expression regulatory elements as well as a recombinant host cell transformed or transfected with the expression vector. The recombinant vector and host cell may be used to express CYP1B1 for protein structure analysis and drug binding studies.

In yet another embodiment, the invention provides a polypeptide comprising a polymorphic variant of a reference amino acid sequence for the CYP1B1 protein. The reference amino acid sequence comprises SEQ ID NO:3 (Fig.3) and the polymorphic variant comprises glycine at a position corresponding to amino acid position 443. In some embodiments, the polymorphic variant also comprises at least one variant amino acid selected from the group consisting of glycine at a position corresponding to amino acid position 48, serine at a position corresponding to amino acid position 119, leucine at a position corresponding to amino acid position 432 and serine at a position corresponding to amino acid position 453. A polymorphic variant of CYP1B1 is useful in studying the effect of the variation on the biological activity of CYP1B1 as well as on the binding affinity of candidate drugs to CYP1B1, or studying the enzymatic properties of such CYP1A2 variants using these candidate drugs as substrates. Herein, the term drug refers to a candidate drug or any of its metabolic derivatives.

The present invention also provides antibodies that recognize and bind to the above polymorphic CYP1B1 protein variant. Such antibodies can be utilized in a variety of diagnostic and prognostic formats and therapeutic methods.

The present invention also provides nonhuman transgenic animals comprising one or more of the CYP1B1 polymorphic genomic variants described herein and methods for producing such animals. The transgenic animals are useful for studying expression of the CYP1B1 isogenes *in vivo*, for *in vivo* screening and testing of drugs targeted against CYP1B1 protein, and for testing the efficacy of therapeutic agents and compounds for breast cancer and primary congenital glaucoma in a biological system.

The present invention also provides a computer system for storing and displaying polymorphism data determined for the CYP1B1 gene. The computer system comprises a computer processing unit; a display; and a database containing the polymorphism data. The polymorphism data includes one or more of the following: the polymorphisms, the genotypes, the haplotypes, and the haplotype pairs identified for the CYP1B1 gene in a reference population. In a preferred embodiment, the computer system is capable of producing a display showing CYP1B1 haplotypes organized

according to their evolutionary relationships.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a reference sequence for the CYP1B1 gene (Genaissance Reference No. 1834513; contiguous lines), with the start and stop positions of each region of coding sequence indicated with a bracket ([or]) and the numerical position below the sequence and the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the sequence. SEQ ID NO:1 is equivalent to Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate nucleotide symbol (R= G or A, Y= T or C, M= A or C, K= G or T, S= G or C, and W= A or T; WIPO standard ST.25). SEQ ID NO:74 is a modified version of SEQ ID NO:1 that shows the context sequence of each polymorphic site, PS1-PS20, in a uniform format to facilitate electronic searching. For each polymorphic site, SEQ ID NO:74 contains a block of 60 bases of the nucleotide sequence encompassing the centrally-located polymorphic site at the 30th position, followed by 60 bases of unspecified sequence to represent that each PS is separated by genomic sequence whose composition is defined elsewhere herein.

Figure 2 illustrates a reference sequence for the CYP1B1 coding sequence (contiguous lines; SEQ ID NO:2), with the polymorphic site(s) and polymorphism(s) identified by Applicants in a reference population indicated by the variant nucleotide positioned below the polymorphic site in the sequence.

Figure 3 illustrates a reference sequence for the CYP1B1 protein (contiguous lines; SEQ ID NO:3), with the variant amino acid(s) caused by the polymorphism(s) of Figure 2 positioned below the polymorphic site in the sequence.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is based on the discovery of novel variants of the CYP1B1 gene. As described in more detail below, the inventors herein discovered 20 isogenes of the CYP1B1 gene by characterizing the CYP1B1 gene found in genomic DNAs isolated from an Index Repository that contains immortalized cell lines from one chimpanzee and 93 human individuals. The human individuals included a reference population of 79 unrelated individuals self-identified as belonging to one of four major population groups: Caucasian (21 individuals), African descent (20 individuals), Asian (20 individuals), or Hispanic/Latino (18 individuals). To the extent possible, the members of this reference population were organized into population subgroups by their self-identified ethnogeographic origin as shown in Table 1 below. In addition, the Index Repository contains three unrelated indigenous American Indians (one from each of North, Central and South America), one three-generation Caucasian family (from the CEPH Utah cohort) and one two-generation African-American family.

Table 1. Population Groups in the Index Repository

Population Group	Population Subgroup	No. of Individuals
African descent		20
	Sierra Leone	1
Asian		20
	Burma	1
	China	3
	Japan	6
	Korea	1
	Philippines	5
	Vietnam	4
Caucasian		21
	British Isles	3
	British Isles/Central	4
	British Isles/Eastern	1
	Central/Eastern	1
	Eastern	3
	Central/Mediterranean	1
	Mediterranean	2
	Scandinavian	2
Hispanic/Latino		18
	Caribbean	8
	Caribbean (Spanish Descent)	2
	Central American (Spanish Descent)	1
	Mexican American	4
	South American (Spanish Descent)	3

The CYP1B1 isogenes present in the human reference population are defined by haplotypes for 20 polymorphic sites in the CYP1B1 gene, 14 of which are believed to be novel. The CYP1B1 polymorphic sites identified by the inventors are referred to as PS1-PS20 to designate the order in which they are located in the gene (see Table 3 below), with the novel polymorphic sites referred to as PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19. Using the genotypes identified in the Index Repository for PS1-PS20 and the methodology described in the Examples below, the inventors herein also determined the pair of haplotypes for the CYP1B1 gene present in individual human members of this repository. The human genotypes and haplotypes found in the repository for the CYP1B1 gene include those shown in Tables 4 and 5, respectively. The polymorphism and haplotype data disclosed herein are useful for validating whether CYP1B1 is a suitable target for drugs to treat breast cancer and primary congenital glaucoma, screening for such drugs and reducing bias in clinical trials of such drugs.

In the context of this disclosure, the following terms shall be defined as follows unless otherwise indicated:

Allele - A particular form of a genetic locus, distinguished from other forms by its particular nucleotide sequence.

Candidate Gene – A gene which is hypothesized to be responsible for a disease, condition, or the response to a treatment, or to be correlated with one of these.

Gene - A segment of DNA that contains all the information for the regulated biosynthesis of an RNA product, including promoters, exons, introns, and other untranslated regions that control expression.

Genotype – An unphased 5′ to 3′ sequence of nucleotide pair(s) found at one or more polymorphic sites in a locus on a pair of homologous chromosomes in an individual. As used herein, genotype includes a full-genotype and/or a sub-genotype as described below.

Full-genotype – The unphased 5′ to 3′ sequence of nucleotide pairs found at all polymorphic sites examined herein in a locus on a pair of homologous chromosomes in a single individual.

Sub-genotype – The unphased 5′ to 3′ sequence of nucleotides seen at a subset of the polymorphic sites examined herein in a locus on a pair of homologous chromosomes in a single individual.

Genotyping – A process for determining a genotype of an individual.

Haplotype – A 5′ to 3′ sequence of nucleotides found at one or more polymorphic sites in a locus on a single chromosome from a single individual. As used herein, haplotype includes a full-haplotype and/or a sub-haplotype as described below.

Full-haplotype – The 5′ to 3′ sequence of nucleotides found at all polymorphic sites examined herein in a locus on a single chromosome from a single individual.

Sub-haplotype – The 5′ to 3′ sequence of nucleotides seen at a subset of the polymorphic sites examined herein in a locus on a single chromosome from a single individual.

Haplotype pair – The two haplotypes found for a locus in a single individual.

Haplotyping – A process for determining one or more haplotypes in an individual and includes use of family pedigrees, molecular techniques and/or statistical inference.

Haplotype data - Information concerning one or more of the following for a specific gene: a listing of the haplotype pairs in each individual in a population; a listing of the different haplotypes in a population; frequency of each haplotype in that or other populations, and any known associations between one or more haplotypes and a trait.

Isoform – A particular form of a gene, mRNA, cDNA, coding sequence or the protein encoded thereby, distinguished from other forms by its particular sequence and/or structure.

Isogene – One of the isoforms (e.g., alleles) of a gene found in a population. An isogene (or allele) contains all of the polymorphisms present in the particular isoform of the gene.

Isolated – As applied to a biological molecule such as RNA, DNA, oligonucleotide, or protein, isolated means the molecule is substantially free of other biological molecules such as nucleic acids, proteins, lipids, carbohydrates, or other material such as cellular debris and growth media. Generally, the term "isolated" is not intended to refer to a complete absence of such material or to absence of water, buffers, or salts, unless they are present in amounts that substantially interfere with the methods of the present invention.

Locus - A location on a chromosome or DNA molecule corresponding to a gene or a physical

or phenotypic feature, where physical features include polymorphic sites.

Naturally-occurring – A term used to designate that the object it is applied to, e.g., naturally-occurring polynucleotide or polypeptide, can be isolated from a source in nature and which has not been intentionally modified by man.

Nucleotide pair – The nucleotides found at a polymorphic site on the two copies of a chromosome from an individual.

Phased – As applied to a sequence of nucleotide pairs for two or more polymorphic sites in a locus, phased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is known.

Polymorphic site (PS) – A position on a chromosome or DNA molecule at which at least two alternative sequences are found in a population.

Polymorphic variant (variant)– A gene, mRNA, cDNA, polypeptide, protein or peptide whose nucleotide or amino acid sequence varies from a reference sequence due to the presence of a polymorphism in the gene.

Polymorphism – The sequence variation observed in an individual at a polymorphic site. Polymorphisms include nucleotide substitutions, insertions, deletions and microsatellites and may, but need not, result in detectable differences in gene expression or protein function.

Polymorphism data – Information concerning one or more of the following for a specific gene: location of polymorphic sites; sequence variation at those sites; frequency of polymorphisms in one or more populations; the different genotypes and/or haplotypes determined for the gene; frequency of one or more of these genotypes and/or haplotypes in one or more populations; any known association(s) between a trait and a genotype or a haplotype for the gene.

Polymorphism Database – A collection of polymorphism data arranged in a systematic or methodical way and capable of being individually accessed by electronic or other means.

Polynucleotide – A nucleic acid molecule comprised of single-stranded RNA or DNA or comprised of complementary, double-stranded DNA.

Population Group – A group of individuals sharing a common ethnogeographic origin.

Reference Population – A group of subjects or individuals who are predicted to be representative of the genetic variation found in the general population. Typically, the reference population represents the genetic variation in the population at a certainty level of at least 85%, preferably at least 90%, more preferably at least 95% and even more preferably at least 99%.

Single Nucleotide Polymorphism (SNP) – Typically, the specific pair of nucleotides observed at a single polymorphic site. In rare cases, three or four nucleotides may be found.

Subject – A human individual whose genotypes or haplotypes or response to treatment or disease state are to be determined.

Treatment - A stimulus administered internally or externally to a subject.

Unphased – As applied to a sequence of nucleotide pairs for two or more polymorphic sites in

a locus, unphased means the combination of nucleotides present at those polymorphic sites on a single copy of the locus is not known.

As discussed above, information on the identity of genotypes and haplotypes for the CYP1B1 gene of any particular individual as well as the frequency of such genotypes and haplotypes in any particular population of individuals is useful for a variety of drug discovery and development applications. Thus, the invention also provides compositions and methods for detecting the novel CYP1B1 polymorphisms, haplotypes and haplotype pairs identified herein.

The compositions comprise at least one oligonucleotide for detecting the variant nucleotide or nucleotide pair located at a novel CYP1B1 polymorphic site in one copy or two copies of the CYP1B1 gene. Such oligonucleotides are referred to herein as CYP1B1 haplotyping oligonucleotides or genotyping oligonucleotides, respectively, and collectively as CYP1B1 oligonucleotides. In one embodiment, a CYP1B1 haplotyping or genotyping oligonucleotide is a probe or primer capable of hybridizing to a target region that contains, or that is located close to, one of the novel polymorphic sites described herein.

As used herein, the term "oligonucleotide" refers to a polynucleotide molecule having less than about 100 nucleotides. A preferred oligonucleotide of the invention is 10 to 35 nucleotides long. More preferably, the oligonucleotide is between 15 and 30, and most preferably, between 20 and 25 nucleotides in length. The exact length of the oligonucleotide will depend on many factors that are routinely considered and practiced by the skilled artisan. The oligonucleotide may be comprised of any phosphorylation state of ribonucleotides, deoxyribonucleotides, and acyclic nucleotide derivatives, and other functionally equivalent derivatives. Alternatively, oligonucleotides may have a phosphate-free backbone, which may be comprised of linkages such as carboxymethyl, acetamidate, carbamate, polyamide (peptide nucleic acid (PNA)) and the like (Varma, R. in Molecular Biology and Biotechnology, A Comprehensive Desk Reference, Ed. R. Meyers, VCH Publishers, Inc. (1995), pages 617-620). Oligonucleotides of the invention may be prepared by chemical synthesis using any suitable methodology known in the art, or may be derived from a biological sample, for example, by restriction digestion. The oligonucleotides may be labeled, according to any technique known in the art, including use of radiolabels, fluorescent labels, enzymatic labels, proteins, haptens, antibodies, sequence tags and the like.

Haplotyping or genotyping oligonucleotides of the invention must be capable of specifically hybridizing to a target region of a CYP1B1 polynucleotide. Preferably, the target region is located in a CYP1B1 isogene. As used herein, specific hybridization means the oligonucleotide forms an anti-parallel double-stranded structure with the target region under certain hybridizing conditions, while failing to form such a structure when incubated with another region in the CYP1B1 polynucleotide or with a non-CYP1B1 polynucleotide under the same hybridizing conditions. Preferably, the oligonucleotide specifically hybridizes to the target region under conventional high stringency conditions. The skilled artisan can readily design and test oligonucleotide probes and primers suitable

for detecting polymorphisms in the CYP1B1 gene using the polymorphism information provided herein in conjunction with the known sequence information for the CYP1B1 gene and routine techniques.

A nucleic acid molecule such as an oligonucleotide or polynucleotide is said to be a "perfect" or "complete" complement of another nucleic acid molecule if every nucleotide of one of the molecules is complementary to the nucleotide at the corresponding position of the other molecule. A nucleic acid molecule is "substantially complementary" to another molecule if it hybridizes to that molecule with sufficient stability to remain in a duplex form under conventional low-stringency conditions. Conventional hybridization conditions are described, for example, by Sambrook J. et al., in Molecular Cloning, A Laboratory Manual, 2nd Edition, Cold Spring Harbor Press, Cold Spring Harbor, NY (1989) and by Haymes, B.D. et al. in Nucleic Acid Hybridization, A Practical Approach, IRL Press, Washington, D.C. (1985). While perfectly complementary oligonucleotides are preferred for detecting polymorphisms, departures from complete complementarity are contemplated where such departures do not prevent the molecule from specifically hybridizing to the target region. For example, an oligonucleotide primer may have a non-complementary fragment at its 5′ end, with the remainder of the primer being complementary to the target region. Alternatively, non-complementary nucleotides may be interspersed into the probe or primer as long as the resulting probe or primer is still capable of specifically hybridizing to the target region.

Preferred haplotyping or genotyping oligonucleotides of the invention are allele-specific oligonucleotides. As used herein, the term allele-specific oligonucleotide (ASO) means an oligonucleotide that is able, under sufficiently stringent conditions, to hybridize specifically to one allele of a gene, or other locus, at a target region containing a polymorphic site while not hybridizing to the corresponding region in another allele(s). As understood by the skilled artisan, allele-specificity will depend upon a variety of readily optimized stringency conditions, including salt and formamide concentrations, as well as temperatures for both the hybridization and washing steps. Examples of hybridization and washing conditions typically used for ASO probes are found in Kogan et al., "Genetic Prediction of Hemophilia A" in PCR Protocols, A Guide to Methods and Applications, Academic Press, 1990 and Ruaño et al., 87 *Proc. Natl. Acad. Sci. USA* 6296-6300, 1990. Typically, an ASO will be perfectly complementary to one allele while containing a single mismatch for another allele.

Allele-specific oligonucleotides of the invention include ASO probes and ASO primers. ASO probes which usually provide good discrimination between different alleles are those in which a central position of the oligonucleotide probe aligns with the polymorphic site in the target region (e.g., approximately the 7th or 8th position in a 15mer, the 8th or 9th position in a 16mer, and the 10th or 11th position in a 20mer). An ASO primer of the invention has a 3′ terminal nucleotide, or preferably a 3′ penultimate nucleotide, that is complementary to only one nucleotide of a particular SNP, thereby acting as a primer for polymerase-mediated extension only if the allele containing that nucleotide is

present. ASO probes and primers hybridizing to either the coding or noncoding strand are contemplated by the invention. ASO probes and primers listed below use the appropriate nucleotide symbol (R= G or A, Y= T or C, M= A or C, K= G or T, S= G or C, and W= A or T; WIPO standard ST.25) at the position of the polymorphic site to represent that the ASO contains either of the two alternative allelic variants observed at that polymorphic site.

A preferred ASO probe for detecting CYP1B1 gene polymorphisms comprises a nucleotide sequence, listed 5′ to 3′, selected from the group consisting of:

CCGCGCCYACCAGCG (SEQ ID NO:4) and its complement,
AGGAGCCYTTGTGTG (SEQ ID NO:5) and its complement,
CTTTCCGRGAAGCAA (SEQ ID NO:6) and its complement,
GCTCAAGYCGCGGAG (SEQ ID NO:7) and its complement,
GCGGCCTYGATTGGA (SEQ ID NO:8) and its complement,
CCTTCTCYTCTCTGT (SEQ ID NO:9) and its complement,
TTCTCCTYTCTGTCC (SEQ ID NO:10) and its complement,
CGGACGGMGCCTTCC (SEQ ID NO:11) and its complement,
TGGACGTSATGCCCT (SEQ ID NO:12) and its complement,
TTCTCCTYTGAAAAA (SEQ ID NO:13) and its complement,
ACAGGTAYCCTGATG (SEQ ID NO:14) and its complement,
TTTATGARGCCATGC (SEQ ID NO:15) and its complement,
GATCCAGSTCGATTC (SEQ ID NO:16) and its complement, and
AATTAGCRTTTAAGG (SEQ ID NO:17) and its complement.

A preferred ASO primer for detecting CYP1B1 gene polymorphisms comprises a nucleotide sequence, listed 5′ to 3′, selected from the group consisting of:

GCGGCCCCGCGCCYA (SEQ ID NO:18); GCCGCCCGCTGGTRG (SEQ ID NO:19);
CGCCCCAGGAGCCYT (SEQ ID NO:20); CTTGGGCACACAARG (SEQ ID NO:21);
CACTGGCTTTCCGRG (SEQ ID NO:22); TTGAGCTTGCTTCYC (SEQ ID NO:23);
AAGCAAGCTCAAGYC (SEQ ID NO:24); TTCCCTCTCCGCGRC (SEQ ID NO:25);
CACCGTGCGGCCTYG (SEQ ID NO:26); GCCACCTCCAATCRA (SEQ ID NO:27);
GTCACGCCTTCTCYT (SEQ ID NO:28); CTGGGGACAGAGARG (SEQ ID NO:29);
CACGCCTTCTCCTYT (SEQ ID NO:30); TGCTGGGGACAGARA (SEQ ID NO:31);
GCAGCGCGGACGGMG (SEQ ID NO:32); GGTCGAGGAAGGCKC (SEQ ID NO:33);
GCCTGGTGGACGTSA (SEQ ID NO:34); GCAGCCAGGGCATSA (SEQ ID NO:35);
GGTCTTTTCTCCTYT (SEQ ID NO:36); TCCGCCTTTTTCARA (SEQ ID NO:37);
CACCAAACAGGTAYC (SEQ ID NO:38); TCTGCACATCAGGRT (SEQ ID NO:39);
CCTTCCTTTATGARG (SEQ ID NO:40); AGAAGCGCATGGCYT (SEQ ID NO:41);
AACTTTGATCCAGST (SEQ ID NO:42); GTCCAAGAATCGASC (SEQ ID NO:43);
CTTCTCAATTAGCRT (SEQ ID NO:44); and TGCTCACCTTAAAYG (SEQ ID NO:45).

Other oligonucleotides of the invention hybridize to a target region located one to several nucleotides downstream of one of the novel polymorphic sites identified herein. Such oligonucleotides are useful in polymerase-mediated primer extension methods for detecting one of the novel polymorphisms described herein and therefore such oligonucleotides are referred to herein as "primer-extension oligonucleotides". In a preferred embodiment, the 3′-terminus of a primer-extension oligonucleotide is a deoxynucleotide complementary to the nucleotide located immediately adjacent to the polymorphic site.

A particularly preferred oligonucleotide primer for detecting CYP1B1 gene polymorphisms

by primer extension terminates in a nucleotide sequence, listed 5′ to 3′, selected from the group consisting of:

GCCCCGCGCC (SEQ ID NO:46); GCCCGCTGGT (SEQ ID NO:47);
CCCAGGAGCC (SEQ ID NO:48); GGGCACACAA (SEQ ID NO:49);
TGGCTTTCCG (SEQ ID NO:50); AGCTTGCTTC (SEQ ID NO:51);
CAAGCTCAAG (SEQ ID NO:52); CCTCTCCGCG (SEQ ID NO:53);
CGTGCGGCCT (SEQ ID NO:54); ACCTCCAATC (SEQ ID NO:55);
ACGCCTTCTC (SEQ ID NO:56); GGGACAGAGA (SEQ ID NO:57);
GCCTTCTCCT (SEQ ID NO:58); TGGGGACAGA (SEQ ID NO:59);
GCGCGGACGG (SEQ ID NO:60); CGAGGAAGGC (SEQ ID NO:61);
TGGTGGACGT (SEQ ID NO:62); GCCAGGGCAT (SEQ ID NO:63);
CTTTTCTCCT (SEQ ID NO:64); GCCTTTTTCA (SEQ ID NO:65);
CAAACAGGTA (SEQ ID NO:66); GCACATCAGG (SEQ ID NO:67);
TCCTTTATGA (SEQ ID NO:68); AGCGCATGGC (SEQ ID NO:69);
TTTGATCCAG (SEQ ID NO:70); CAAGAATCGA (SEQ ID NO:71);
CTCAATTAGC (SEQ ID NO:72); and TCACCTTAAA (SEQ ID NO:73).

In some embodiments, a composition contains two or more differently labeled CYP1B1 oligonucleotides for simultaneously probing the identity of nucleotides or nucleotide pairs at two or more polymorphic sites. It is also contemplated that primer compositions may contain two or more sets of allele-specific primer pairs to allow simultaneous targeting and amplification of two or more regions containing a polymorphic site.

CYP1B1 oligonucleotides of the invention may also be immobilized on or synthesized on a solid surface such as a microchip, bead, or glass slide (see, e.g., WO 98/20020 and WO 98/20019). Such immobilized oligonucleotides may be used in a variety of polymorphism detection assays, including but not limited to probe hybridization and polymerase extension assays. Immobilized CYP1B1 oligonucleotides of the invention may comprise an ordered array of oligonucleotides designed to rapidly screen a DNA sample for polymorphisms in multiple genes at the same time.

In another embodiment, the invention provides a kit comprising at least two CYP1B1 oligonucleotides packaged in separate containers. The kit may also contain other components such as hybridization buffer (where the oligonucleotides are to be used as a probe) packaged in a separate container. Alternatively, where the oligonucleotides are to be used to amplify a target region, the kit may contain, packaged in separate containers, a polymerase and a reaction buffer optimized for primer extension mediated by the polymerase, such as PCR.

The above described oligonucleotide compositions and kits are useful in methods for genotyping and/or haplotyping the CYP1B1 gene in an individual. As used herein, the terms "CYP1B1 genotype" and "CYP1B1 haplotype" mean the genotype or haplotype contains the nucleotide pair or nucleotide, respectively, that is present at one or more of the novel polymorphic sites described herein and may optionally also include the nucleotide pair or nucleotide present at one or more additional polymorphic sites in the CYP1B1 gene. The additional polymorphic sites may be currently known polymorphic sites or sites that are subsequently discovered.

One embodiment of a genotyping method of the invention involves isolating from the

individual a nucleic acid sample comprising the two copies of the CYP1B1 gene, mRNA transcripts thereof or cDNA copies thereof, or a fragment of any of the foregoing, that are present in the individual, and determining the identity of the nucleotide pair at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19 in the two copies to assign a CYP1B1 genotype to the individual. As will be readily understood by the skilled artisan, the two "copies" of a gene, mRNA or cDNA (or fragment of such CYP1B1 molecules) in an individual may be the same allele or may be different alleles. In a preferred embodiment of the method for assigning a CYP1B1 genotype, the identity of the nucleotide pair at one or more of the polymorphic sites selected from the group consisting of PS8, PS9, PS15, PS17, PS18 and PS20 is also determined. In another embodiment, a genotyping method of the invention comprises determining the identity of the nucleotide pair at each of PS1-PS20.

Typically, the nucleic acid sample is isolated from a biological sample taken from the individual, such as a blood sample or tissue sample. Suitable tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. The nucleic acid sample may be comprised of genomic DNA, mRNA, or cDNA and, in the latter two cases, the biological sample must be obtained from a tissue in which the CYP1B1 gene is expressed. Furthermore it will be understood by the skilled artisan that mRNA or cDNA preparations would not be used to detect polymorphisms located in introns or in 5′ and 3′ untranslated regions if not present in the mRNA or cDNA. If a CYP1B1 gene fragment is isolated, it must contain the polymorphic site(s) to be genotyped.

One embodiment of a haplotyping method of the invention comprises isolating from the individual a nucleic acid sample containing only one of the two copies of the CYP1B1 gene, mRNA or cDNA, or a fragment of such CYP1B1 molecules, that is present in the individual and determining in that copy the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19 in that copy to assign a CYP1B1 haplotype to the individual.

The nucleic acid used in the above haplotyping methods of the invention may be isolated using any method capable of separating the two copies of the CYP1B1 gene or fragment such as one of the methods described above for preparing CYP1B1 isogenes, with targeted *in vivo* cloning being the preferred approach. As will be readily appreciated by those skilled in the art, any individual clone will typically only provide haplotype information on one of the two CYP1B1 gene copies present in an individual. If haplotype information is desired for the individual's other copy, additional CYP1B1 clones will usually need to be examined. Typically, at least five clones should be examined to have more than a 90% probability of haplotyping both copies of the CYP1B1 gene in an individual. In some cases, however, once the haplotype for one CYP1B1 allele is directly determined, the haplotype for the other allele may be inferred if the individual has a known genotype for the polymorphic sites of interest or if the haplotype frequency or haplotype pair frequency for the individual's population group

is known. In some embodiments, the CYP1B1 haplotype is assigned to the individual by also identifying the nucleotide at one or more polymorphic sites selected from the group consisting of PS8, PS9, PS15, PS17, PS18 and PS20. In a particularly preferred embodiment, the nucleotide at each of PS1-PS20 is identified.

In another embodiment, the haplotyping method comprises determining whether an individual has one or more of the CYP1B1 haplotypes shown in Table 5. This can be accomplished by identifying, for one or both copies of the individual's CYP1B1 gene, the phased sequence of nucleotides present at each of PS1-PS20. This identifying step does not necessarily require that each of PS1-PS20 be directly examined. Typically only a subset of PS1-PS20 will need to be directly examined to assign to an individual one or more of the haplotypes shown in Table 5. This is because at least one polymorphic site in a gene is frequently in strong linkage disequilibrium with one or more other polymorphic sites in that gene (Drysdale, CM et al. 2000 *PNAS* 97:10483-10488; Rieder MJ et al. 1999 *Nature Genetics* 22:59-62). Two sites are said to be in linkage disequilibrium if the presence of a particular variant at one site enhances the predictability of another variant at the second site (Stephens, JC 1999, *Mol. Diag.* 4:309-317). Techniques for determining whether any two polymorphic sites are in linkage disequilibrium are well-known in the art (Weir B.S. 1996 *Genetic Data Analysis II*, Sinauer Associates, Inc. Publishers, Sunderland, MA).

In another embodiment of a haplotyping method of the invention, a CYP1B1 haplotype pair is determined for an individual by identifying the phased sequence of nucleotides at one or more polymorphic sites selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19 in each copy of the CYP1B1 gene that is present in the individual. In a particularly preferred embodiment, the haplotyping method comprises identifying the phased sequence of nucleotides at each of PS1-PS20 in each copy of the CYP1B1 gene.

When haplotyping both copies of the gene, the identifying step is preferably performed with each copy of the gene being placed in separate containers. However, it is also envisioned that if the two copies are labeled with different tags, or are otherwise separately distinguishable or identifiable, it could be possible in some cases to perform the method in the same container. For example, if first and second copies of the gene are labeled with different first and second fluorescent dyes, respectively, and an allele-specific oligonucleotide labeled with yet a third different fluorescent dye is used to assay the polymorphic site(s), then detecting a combination of the first and third dyes would identify the polymorphism in the first gene copy while detecting a combination of the second and third dyes would identify the polymorphism in the second gene copy.

In both the genotyping and haplotyping methods, the identity of a nucleotide (or nucleotide pair) at a polymorphic site(s) may be determined by amplifying a target region(s) containing the polymorphic site(s) directly from one or both copies of the CYP1B1 gene, or a fragment thereof, and the sequence of the amplified region(s) determined by conventional methods. It will be readily appreciated by the skilled artisan that only one nucleotide will be detected at a polymorphic site in

individuals who are homozygous at that site, while two different nucleotides will be detected if the individual is heterozygous for that site. The polymorphism may be identified directly, known as positive-type identification, or by inference, referred to as negative-type identification. For example, where a SNP is known to be guanine and cytosine in a reference population, a site may be positively determined to be either guanine or cytosine for an individual homozygous at that site, or both guanine and cytosine, if the individual is heterozygous at that site. Alternatively, the site may be negatively determined to be not guanine (and thus cytosine/cytosine) or not cytosine (and thus guanine/guanine).

The target region(s) may be amplified using any oligonucleotide-directed amplification method, including but not limited to polymerase chain reaction (PCR) (U.S. Patent No. 4,965,188), ligase chain reaction (LCR) (Barany et al., *Proc. Natl. Acad. Sci. USA* 88:189-193, 1991; WO90/01069), and oligonucleotide ligation assay (OLA) (Landegren et al., *Science* 241:1077-1080, 1988). Other known nucleic acid amplification procedures may be used to amplify the target region including transcription-based amplification systems (U.S. Patent No. 5,130,238; EP 329,822; U.S. Patent No. 5,169,766, WO89/06700) and isothermal methods (Walker et al., *Proc. Natl. Acad. Sci. USA* 89:392-396, 1992).

A polymorphism in the target region may also be assayed before or after amplification using one of several hybridization-based methods known in the art. Typically, allele-specific oligonucleotides are utilized in performing such methods. The allele-specific oligonucleotides may be used as differently labeled probe pairs, with one member of the pair showing a perfect match to one variant of a target sequence and the other member showing a perfect match to a different variant. In some embodiments, more than one polymorphic site may be detected at once using a set of allele-specific oligonucleotides or oligonucleotide pairs. Preferably, the members of the set have melting temperatures within 5°C, and more preferably within 2°C, of each other when hybridizing to each of the polymorphic sites being detected.

Hybridization of an allele-specific oligonucleotide to a target polynucleotide may be performed with both entities in solution, or such hybridization may be performed when either the oligonucleotide or the target polynucleotide is covalently or noncovalently affixed to a solid support. Attachment may be mediated, for example, by antibody-antigen interactions, poly-L-Lys, streptavidin or avidin-biotin, salt bridges, hydrophobic interactions, chemical linkages, UV cross-linking baking, etc. Allele-specific oligonucleotides may be synthesized directly on the solid support or attached to the solid support subsequent to synthesis. Solid-supports suitable for use in detection methods of the invention include substrates made of silicon, glass, plastic, paper and the like, which may be formed, for example, into wells (as in 96-well plates), slides, sheets, membranes, fibers, chips, dishes, and beads. The solid support may be treated, coated or derivatized to facilitate the immobilization of the allele-specific oligonucleotide or target nucleic acid.

The genotype or haplotype for the CYP1B1 gene of an individual may also be determined by hybridization of a nucleic acid sample containing one or both copies of the gene, mRNA, cDNA or

fragment(s) thereof, to nucleic acid arrays and subarrays such as described in WO 95/11995. The arrays would contain a battery of allele-specific oligonucleotides representing each of the polymorphic sites to be included in the genotype or haplotype.

The identity of polymorphisms may also be determined using a mismatch detection technique, including but not limited to the RNase protection method using riboprobes (Winter et al., *Proc. Natl. Acad. Sci. USA* 82:7575, 1985; Meyers et al., *Science* 230:1242, 1985) and proteins which recognize nucleotide mismatches, such as the E. coli mutS protein (Modrich, P. *Ann. Rev. Genet.* 25:229-253, 1991). Alternatively, variant alleles can be identified by single strand conformation polymorphism (SSCP) analysis (Orita et al., *Genomics* 5:874-879, 1989; Humphries et al., in Molecular Diagnosis of Genetic Diseases, R. Elles, ed., pp. 321-340, 1996) or denaturing gradient gel electrophoresis (DGGE) (Wartell et al., *Nucl. Acids Res.* 18:2699-2706, 1990; Sheffield et al., *Proc. Natl. Acad. Sci. USA* 86:232-236, 1989).

A polymerase-mediated primer extension method may also be used to identify the polymorphism(s). Several such methods have been described in the patent and scientific literature and include the "Genetic Bit Analysis" method (WO92/15712) and the ligase/polymerase mediated genetic bit analysis (U.S. Patent 5,679,524. Related methods are disclosed in WO91/02087, WO90/09455, WO95/17676, U.S. Patent Nos. 5,302,509, and 5,945,283. Extended primers containing a polymorphism may be detected by mass spectrometry as described in U.S. Patent No. 5,605,798. Another primer extension method is allele-specific PCR (Ruaño et al., *Nucl. Acids Res.* 17:8392, 1989; Ruaño et al., *Nucl. Acids Res.* 19, 6877-6882, 1991; WO 93/22456; Turki et al., *J. Clin. Invest.* 95:1635-1641, 1995). In addition, multiple polymorphic sites may be investigated by simultaneously amplifying multiple regions of the nucleic acid using sets of allele-specific primers as described in Wallace et al. (WO89/10414).

In addition, the identity of the allele(s) present at any of the novel polymorphic sites described herein may be indirectly determined by haplotyping or genotyping another polymorphic site that is in linkage disequilibrium with the polymorphic site that is of interest. Polymorphic sites in linkage disequilibrium with the presently disclosed polymorphic sites may be located in regions of the gene or in other genomic regions not examined herein. Detection of the allele(s) present at a polymorphic site in linkage disequilibrium with the novel polymorphic sites described herein may be performed by, but is not limited to, any of the above-mentioned methods for detecting the identity of the allele at a polymorphic site.

In another aspect of the invention, an individual's CYP1B1 haplotype pair is predicted from its CYP1B1 genotype using information on haplotype pairs known to exist in a reference population. In its broadest embodiment, the haplotyping prediction method comprises identifying a CYP1B1 genotype for the individual at two or more CYP1B1 polymorphic sites described herein, accessing data containing CYP1B1 haplotype pairs identified in a reference population, and assigning a haplotype pair to the individual that is consistent with the genotype data. In one embodiment, the

reference haplotype pairs include the CYP1B1 haplotype pairs shown in Table 4. The CYP1B1 haplotype pair can be assigned by comparing the individual's genotype with the genotypes corresponding to the haplotype pairs known to exist in the general population or in a specific population group, and determining which haplotype pair is consistent with the genotype of the individual. In some embodiments, the comparing step may be performed by visual inspection (for example, by consulting Table 4). When the genotype of the individual is consistent with more than one haplotype pair, frequency data (such as that presented in Table 7) may be used to determine which of these haplotype pairs is most likely to be present in the individual. This determination may also be performed in some embodiments by visual inspection, for example by consulting Table 7. If a particular CYP1B1 haplotype pair consistent with the genotype of the individual is more frequent in the reference population than others consistent with the genotype, then that haplotype pair with the highest frequency is the most likely to be present in the individual. In other embodiments, the comparison may be made by a computer-implemented algorithm with the genotype of the individual and the reference haplotype data stored in computer-readable formats. For example, as described in PCT/US01/12831, filed April 18, 2001, one computer-implemented algorithm to perform this comparison entails enumerating all possible haplotype pairs which are consistent with the genotype, accessing data containing CYP1B1 haplotype pairs frequency data determined in a reference population to determine a probability that the individual has a possible haplotype pair, and analyzing the determined probabilities to assign a haplotype pair to the individual.

Generally, the reference population should be composed of randomly-selected individuals representing the major ethnogeographic groups of the world. A preferred reference population for use in the methods of the present invention comprises an approximately equal number of individuals from Caucasian, African-descent, Asian and Hispanic-Latino population groups with the minimum number of each group being chosen based on how rare a haplotype one wants to be guaranteed to see. For example, if one wants to have a q% chance of not missing a haplotype that exists in the population at a p% frequency of occurring in the reference population, the number of individuals (n) who must be sampled is given by $2n=\log(1-q)/\log(1-p)$ where p and q are expressed as fractions. A preferred reference population allows the detection of any haplotype whose frequency is at least 10% with about 99% certainty and comprises about 20 unrelated individuals from each of the four population groups named above. A particularly preferred reference population includes a 3-generation family representing one or more of the four population groups to serve as controls for checking quality of haplotyping procedures.

In a preferred embodiment, the haplotype frequency data for each ethnogeographic group is examined to determine whether it is consistent with Hardy-Weinberg equilibrium. Hardy-Weinberg equilibrium (D.L. Hartl et al., Principles of Population Genomics, Sinauer Associates (Sunderland, MA), 3rd Ed., 1997) postulates that the frequency of finding the haplotype pair H_1 / H_2 is equal to

$p_{H-W}(H_1 / H_2) = 2p(H_1)p(H_2)$ if $H_1 \neq H_2$ and $p_{H-W}(H_1 / H_2) = p(H_1)p(H_2)$ if $H_1 = H_2$.

A statistically significant difference between the observed and expected haplotype frequencies could be due to one or more factors including significant inbreeding in the population group, strong selective pressure on the gene, sampling bias, and/or errors in the genotyping process. If large deviations from Hardy-Weinberg equilibrium are observed in an ethnogeographic group, the number of individuals in that group can be increased to see if the deviation is due to a sampling bias. If a larger sample size does not reduce the difference between observed and expected haplotype pair frequencies, then one may wish to consider haplotyping the individual using a direct haplotyping method such as, for example, CLASPER System™ technology (U.S. Patent No. 5,866,404), single molecule dilution, or allele-specific long-range PCR (Michalotos-Beloin et al., *Nucleic Acids Res.* 24:4841-4843, 1996).

In one embodiment of this method for predicting a CYP1B1 haplotype pair for an individual, the assigning step involves performing the following analysis. First, each of the possible haplotype pairs is compared to the haplotype pairs in the reference population. Generally, only one of the haplotype pairs in the reference population matches a possible haplotype pair and that pair is assigned to the individual. Occasionally, only one haplotype represented in the reference haplotype pairs is consistent with a possible haplotype pair for an individual, and in such cases the individual is assigned a haplotype pair containing this known haplotype and a new haplotype derived by subtracting the known haplotype from the possible haplotype pair. Alternatively, the haplotype pair in an individual may be predicted from the individual's genotype for that gene using reported methods (e.g., Clark et al. 1990 *Mol Bio Evol* 7:111-22; copending PCT/US01/12831 filed April 18, 2001) or through a commercial haplotyping service such as offered by Genaissance Pharmaceuticals, Inc. (New Haven, CT). In rare cases, either no haplotypes in the reference population are consistent with the possible haplotype pairs, or alternatively, multiple reference haplotype pairs are consistent with the possible haplotype pairs. In such cases, the individual is preferably haplotyped using a direct molecular haplotyping method such as, for example, CLASPER System™ technology (U.S. Patent No. 5,866,404), SMD, or allele-specific long-range PCR (Michalotos-Beloin et al., *supra*).

The invention also provides a method for determining the frequency of a CYP1B1 genotype, haplotype, or haplotype pair in a population. The method comprises, for each member of the population, determining the genotype or the haplotype pair for the novel CYP1B1 polymorphic sites described herein, and calculating the frequency any particular genotype, haplotype, or haplotype pair is found in the population. The population may be e.g., a reference population, a family population, a same gender population, a population group, or a trait population (e.g., a group of individuals exhibiting a trait of interest such as a medical condition or response to a therapeutic treatment).

In another aspect of the invention, frequency data for CYP1B1 genotypes, haplotypes, and/or haplotype pairs are determined in a reference population and used in a method for identifying an association between a trait and a CYP1B1 genotype, haplotype, or haplotype pair. The trait may be

any detectable phenotype, including but not limited to susceptibility to a disease or response to a treatment. In one embodiment, the method involves obtaining data on the frequency of the genotype(s), haplotype(s), or haplotype pair(s) of interest in a reference population as well as in a population exhibiting the trait. Frequency data for one or both of the reference and trait populations may be obtained by genotyping or haplotyping each individual in the populations using one or more of the methods described above. The haplotypes for the trait population may be determined directly or, alternatively, by a predictive genotype to haplotype approach as described above. In another embodiment, the frequency data for the reference and/or trait populations is obtained by accessing previously determined frequency data, which may be in written or electronic form. For example, the frequency data may be present in a database that is accessible by a computer. Once the frequency data is obtained, the frequencies of the genotype(s), haplotype(s), or haplotype pair(s) of interest in the reference and trait populations are compared. In a preferred embodiment, the frequencies of all genotypes, haplotypes, and/or haplotype pairs observed in the populations are compared. If a particular CYP1B1 genotype, haplotype, or haplotype pair is more frequent in the trait population than in the reference population at a statistically significant amount, then the trait is predicted to be associated with that CYP1B1 genotype, haplotype or haplotype pair. Preferably, the CYP1B1 genotype, haplotype, or haplotype pair being compared in the trait and reference populations is selected from the full-genotypes and full-haplotypes shown in Tables 4 and 5, or from sub-genotypes and sub-haplotypes derived from these genotypes and haplotypes. Sub-genotypes useful in the invention preferably do not include sub-genotypes solely for any one of PS8, PS9, PS15, PS17, PS18 and PS20 or for any combination thereof.

In a preferred embodiment of the method, the trait of interest is a clinical response exhibited by a patient to some therapeutic treatment, for example, response to a drug targeting CYP1B1 or response to a therapeutic treatment for a medical condition. As used herein, "medical condition" includes but is not limited to any condition or disease manifested as one or more physical and/or psychological symptoms for which treatment is desirable, and includes previously and newly identified diseases and other disorders. As used herein the term "clinical response" means any or all of the following: a quantitative measure of the response, no response, and/or adverse response (i.e., side effects).

In order to deduce a correlation between clinical response to a treatment and a CYP1B1 genotype, haplotype, or haplotype pair, it is necessary to obtain data on the clinical responses exhibited by a population of individuals who received the treatment, hereinafter the "clinical population". This clinical data may be obtained by analyzing the results of a clinical trial that has already been run and/or the clinical data may be obtained by designing and carrying out one or more new clinical trials. As used herein, the term "clinical trial" means any research study designed to collect clinical data on responses to a particular treatment, and includes but is not limited to phase I, phase II and phase III clinical trials. Standard methods are used to define the patient population and to

enroll subjects.

It is preferred that the individuals included in the clinical population have been graded for the existence of the medical condition of interest. This is important in cases where the symptom(s) being presented by the patients can be caused by more than one underlying condition, and where treatment of the underlying conditions are not the same. An example of this would be where patients experience breathing difficulties that are due to either asthma or respiratory infections. If both sets were treated with an asthma medication, there would be a spurious group of apparent non-responders that did not actually have asthma. These people would affect the ability to detect any correlation between haplotype and treatment outcome. This grading of potential patients could employ a standard physical exam or one or more lab tests. Alternatively, grading of patients could use haplotyping for situations where there is a strong correlation between haplotype pair and disease susceptibility or severity.

The therapeutic treatment of interest is administered to each individual in the trial population and each individual's response to the treatment is measured using one or more predetermined criteria. It is contemplated that in many cases, the trial population will exhibit a range of responses and that the investigator will choose the number of responder groups (e.g., low, medium, high) made up by the various responses. In addition, the CYP1B1 gene for each individual in the trial population is genotyped and/or haplotyped, which may be done before or after administering the treatment.

After both the clinical and polymorphism data have been obtained, correlations between individual response and CYP1B1 genotype or haplotype content are created. Correlations may be produced in several ways. In one method, individuals are grouped by their CYP1B1 genotype or haplotype (or haplotype pair) (also referred to as a polymorphism group), and then the averages and standard deviations of clinical responses exhibited by the members of each polymorphism group are calculated.

These results are then analyzed to determine if any observed variation in clinical response between polymorphism groups is statistically significant. Statistical analysis methods which may be used are described in L.D. Fisher and G. vanBelle, "Biostatistics: A Methodology for the Health Sciences", Wiley-Interscience (New York) 1993. This analysis may also include a regression calculation of which polymorphic sites in the CYP1B1 gene give the most significant contribution to the differences in phenotype. One regression model useful in the invention is described in WO 01/01218, entitled "Methods for Obtaining and Using Haplotype Data".

A second method for finding correlations between CYP1B1 haplotype content and clinical responses uses predictive models based on error-minimizing optimization algorithms. One of many possible optimization algorithms is a genetic algorithm (R. Judson, "Genetic Algorithms and Their Uses in Chemistry" in Reviews in Computational Chemistry, Vol. 10, pp. 1-73, K. B. Lipkowitz and D. B. Boyd, eds. (VCH Publishers, New York, 1997). Simulated annealing (Press et al., "Numerical Recipes in C: The Art of Scientific Computing", Cambridge University Press (Cambridge) 1992, Ch. 10), neural networks (E. Rich and K. Knight, "Artificial Intelligence", 2nd Edition (McGraw-Hill, New

York, 1991, Ch. 18), standard gradient descent methods (Press et al., *supra,* Ch. 10), or other global or local optimization approaches (see discussion in Judson, *supra*) could also be used. Preferably, the correlation is found using a genetic algorithm approach as described in WO 01/01218.

Correlations may also be analyzed using analysis of variation (ANOVA) techniques to determine how much of the variation in the clinical data is explained by different subsets of the polymorphic sites in the CYP1B1 gene. As described in WO 01/01218, ANOVA is used to test hypotheses about whether a response variable is caused by or correlated with one or more traits or variables that can be measured (Fisher and vanBelle, *supra*, Ch. 10).

From the analyses described above, a mathematical model may be readily constructed by the skilled artisan that predicts clinical response as a function of CYP1B1 genotype or haplotype content. Preferably, the model is validated in one or more follow-up clinical trials designed to test the model.

The identification of an association between a clinical response and a genotype or haplotype (or haplotype pair) for the CYP1B1 gene may be the basis for designing a diagnostic method to determine those individuals who will or will not respond to the treatment, or alternatively, will respond at a lower level and thus may require more treatment, i.e., a greater dose of a drug. The diagnostic method may take one of several forms: for example, a direct DNA test (i.e., genotyping or haplotyping one or more of the polymorphic sites in the CYP1B1 gene), a serological test, or a physical exam measurement. The only requirement is that there be a good correlation between the diagnostic test results and the underlying CYP1B1 genotype or haplotype that is in turn correlated with the clinical response. In a preferred embodiment, this diagnostic method uses the predictive haplotyping method described above.

In another embodiment, the invention provides an isolated polynucleotide comprising a polymorphic variant of the CYP1B1 gene or a fragment of the gene which contains at least one of the novel polymorphic sites described herein. The nucleotide sequence of a variant CYP1B1 gene is identical to the reference genomic sequence for those portions of the gene examined, as described in the Examples below, except that it comprises a different nucleotide at one or more of the novel polymorphic sites PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19, and may also comprise one or more additional polymorphisms selected from the group consisting of guanine at PS8, thymine at PS9, cytosine at PS15, cytosine at PS17, guanine at PS18 and guanine at PS20. Similarly, the nucleotide sequence of a variant fragment of the CYP1B1 gene is identical to the corresponding portion of the reference sequence except for having a different nucleotide at one or more of the novel polymorphic sites described herein. Thus, the invention specifically does not include polynucleotides comprising a nucleotide sequence identical to the reference sequence of the CYP1B1 gene, which is defined by haplotype 12, (or other reported CYP1B1 sequences) or to portions of the reference sequence (or other reported CYP1B1 sequences), except for the haplotyping and genotyping oligonucleotides described above.

The location of a polymorphism in a variant CYP1B1 gene or fragment is preferably

identified by aligning its sequence against SEQ ID NO:1. The polymorphism is selected from the group consisting of thymine at PS1, cytosine at PS2, adenine at PS3, cytosine at PS4, thymine at PS5, thymine at PS6, thymine at PS7, adenine at PS10, cytosine at PS11, thymine at PS12, cytosine at PS13, guanine at PS14, guanine at PS16 and adenine at PS19. In a preferred embodiment, the polymorphic variant comprises a naturally-occurring isogene of the CYP1B1 gene which is defined by any one of haplotypes 1- 11 and 13 - 20 shown in Table 5 below.

Polymorphic variants of the invention may be prepared by isolating a clone containing the CYP1B1 gene from a human genomic library. The clone may be sequenced to determine the identity of the nucleotides at the novel polymorphic sites described herein. Any particular variant or fragment thereof, that is claimed herein could be prepared from this clone by performing *in vitro* mutagenesis using procedures well-known in the art. Any particular CYP1B1 variant or fragment thereof may also be prepared using synthetic or semi-synthetic methods known in the art.

CYP1B1 isogenes, or fragments thereof, may be isolated using any method that allows separation of the two "copies" of the CYP1B1 gene present in an individual, which, as readily understood by the skilled artisan, may be the same allele or different alleles. Separation methods include targeted *in vivo* cloning (TIVC) in yeast as described in WO 98/01573, U.S. Patent No. 5,866,404, and U.S. Patent No. 5,972,614. Another method, which is described in U.S. Patent No. 5,972,614, uses an allele specific oligonucleotide in combination with primer extension and exonuclease degradation to generate hemizygous DNA targets. Yet other methods are single molecule dilution (SMD) as described in Ruaño et al., *Proc. Natl. Acad. Sci.* 87:6296-6300, 1990; and allele specific PCR (Ruaño et al., 1989, *supra*; Ruaño et al., 1991, *supra*; Michalatos-Beloin et al., *supra*).

The invention also provides CYP1B1 genome anthologies, which are collections of at least two CYP1B1 isogenes found in a given population. The population may be any group of at least two individuals, including but not limited to a reference population, a population group, a family population, a clinical population, and a same gender population. A CYP1B1 genome anthology may comprise individual CYP1B1 isogenes stored in separate containers such as microtest tubes, separate wells of a microtitre plate and the like. Alternatively, two or more groups of the CYP1B1 isogenes in the anthology may be stored in separate containers. Individual isogenes or groups of such isogenes in a genome anthology may be stored in any convenient and stable form, including but not limited to in buffered solutions, as DNA precipitates, freeze-dried preparations and the like. A preferred CYP1B1 genome anthology of the invention comprises a set of isogenes defined by the haplotypes shown in Table 5 below.

An isolated polynucleotide containing a polymorphic variant nucleotide sequence of the invention may be operably linked to one or more expression regulatory elements in a recombinant expression vector capable of being propagated and expressing the encoded CYP1B1 protein in a prokaryotic or a eukaryotic host cell. Examples of expression regulatory elements which may be used include, but are not limited to, the lac system, operator and promoter regions of phage lambda, yeast

promoters, and promoters derived from vaccinia virus, adenovirus, retroviruses, or SV40. Other regulatory elements include, but are not limited to, appropriate leader sequences, termination codons, polyadenylation signals, and other sequences required for the appropriate transcription and subsequent translation of the nucleic acid sequence in a given host cell. Of course, the correct combinations of expression regulatory elements will depend on the host system used. In addition, it is understood that the expression vector contains any additional elements necessary for its transfer to and subsequent replication in the host cell. Examples of such elements include, but are not limited to, origins of replication and selectable markers. Such expression vectors are commercially available or are readily constructed using methods known to those in the art (e.g., F. Ausubel et al., 1987, in "Current Protocols in Molecular Biology", John Wiley and Sons, New York, New York). Host cells which may be used to express the variant CYP1B1 sequences of the invention include, but are not limited to, eukaryotic and mammalian cells, such as animal, plant, insect and yeast cells, and prokaryotic cells, such as E. coli, or algal cells as known in the art. The recombinant expression vector may be introduced into the host cell using any method known to those in the art including, but not limited to, microinjection, electroporation, particle bombardment, transduction, and transfection using DEAE-dextran, lipofection, or calcium phosphate (see e.g., Sambrook et al. (1989) in "Molecular Cloning. A Laboratory Manual", Cold Spring Harbor Press, Plainview, New York). In a preferred aspect, eukaryotic expression vectors that function in eukaryotic cells, and preferably mammalian cells, are used. Non-limiting examples of such vectors include vaccinia virus vectors, adenovirus vectors, herpes virus vectors, and baculovirus transfer vectors. Preferred eukaryotic cell lines include COS cells, CHO cells, HeLa cells, NIH/3T3 cells, and embryonic stem cells (Thomson, J. A. et al., 1998 *Science* 282:1145-1147). Particularly preferred host cells are mammalian cells.

As will be readily recognized by the skilled artisan, expression of polymorphic variants of the CYP1B1 gene will produce CYP1B1 mRNAs varying from each other at any polymorphic site retained in the spliced and processed mRNA molecules. These mRNAs can be used for the preparation of a CYP1B1 cDNA comprising a nucleotide sequence which is a polymorphic variant of the CYP1B1 reference coding sequence shown in Figure 2. Thus, the invention also provides CYP1B1 mRNAs and corresponding cDNAs which comprise a nucleotide sequence that is identical to SEQ ID NO:2 (Fig. 2) (or its corresponding RNA sequence) for those regions of SEQ ID NO:2 that correspond to the examined portions of the CYP1B1 gene (as described in the Examples below), except for having one or more polymorphisms selected from the group consisting of adenine at a position corresponding to nucleotide 564, cytosine at a position corresponding to nucleotide 729, cytosine at a position corresponding to nucleotide 1047, guanine at a position corresponding to nucleotide 1161 and guanine at a position corresponding to nucleotide 1328, and may also comprise one or more additional polymorphisms selected from the group consisting of guanine at a position corresponding to nucleotide 142, thymine at a position corresponding to nucleotide 355, cytosine at a position corresponding to nucleotide 1294, cytosine at a position corresponding to nucleotide 1347

and guanine at a position corresponding to nucleotide 1358. A particularly preferred polymorphic cDNA variant comprises the coding sequence of a CYP1B1 isogene defined by any one of haplotypes 1, 3-11 and 13-20. Fragments of these variant mRNAs and cDNAs are included in the scope of the invention, provided they contain one or more of the novel polymorphisms described herein. The invention specifically excludes polynucleotides identical to previously identified CYP1B1 mRNAs or cDNAs, and previously described fragments thereof. Polynucleotides comprising a variant CYP1B1 RNA or DNA sequence may be isolated from a biological sample using well-known molecular biological procedures or may be chemically synthesized.

As used herein, a polymorphic variant of a CYP1B1 gene, mRNA or cDNA fragment comprises at least one novel polymorphism identified herein and has a length of at least 10 nucleotides and may range up to the full length of the gene. Preferably, such fragments are between 100 and 3000 nucleotides in length, and more preferably between 200 and 2000 nucleotides in length, and most preferably between 500 and 1000 nucleotides in length.

In describing the CYP1B1 polymorphic sites identified herein, reference is made to the sense strand of the gene for convenience. However, as recognized by the skilled artisan, nucleic acid molecules containing the CYP1B1 gene or cDNA may be complementary double stranded molecules and thus reference to a particular site on the sense strand refers as well to the corresponding site on the complementary antisense strand. Thus, reference may be made to the same polymorphic site on either strand and an oligonucleotide may be designed to hybridize specifically to either strand at a target region containing the polymorphic site. Thus, the invention also includes single-stranded polynucleotides which are complementary to the sense strand of the CYP1B1 genomic, mRNA and cDNA variants described herein.

Polynucleotides comprising a polymorphic gene variant or fragment of the invention may be useful for therapeutic purposes. For example, where a patient could benefit from expression, or increased expression, of a particular CYP1B1 protein isoform, an expression vector encoding the isoform may be administered to the patient. The patient may be one who lacks the CYP1B1 isogene encoding that isoform or may already have at least one copy of that isogene.

In other situations, it may be desirable to decrease or block expression of a particular CYP1B1 isogene. Expression of a CYP1B1 isogene may be turned off by transforming a targeted organ, tissue or cell population with an expression vector that expresses high levels of untranslatable mRNA or antisense RNA for the isogene or fragment thereof. Alternatively, oligonucleotides directed against the regulatory regions (e.g., promoter, introns, enhancers, 3′ untranslated region) of the isogene may block transcription. Oligonucleotides targeting the transcription initiation site, e.g., between positions −10 and +10 from the start site are preferred. Similarly, inhibition of transcription can be achieved using oligonucleotides that base-pair with region(s) of the isogene DNA to form triplex DNA (see e.g., Gee et al. in Huber, B.E. and B.I. Carr, Molecular and Immunologic Approaches, Futura Publishing Co., Mt. Kisco, N.Y., 1994). Antisense oligonucleotides may also be designed to block translation of

CYP1B1 mRNA transcribed from a particular isogene. It is also contemplated that ribozymes may be designed that can catalyze the specific cleavage of CYP1B1 mRNA transcribed from a particular isogene.

The untranslated mRNA, antisense RNA or antisense oligonucleotides may be delivered to a target cell or tissue by expression from a vector introduced into the cell or tissue *in vivo* or *ex vivo*. Alternatively, such molecules may be formulated as a pharmaceutical composition for administration to the patient. Oligoribonucleotides and/or oligodeoxynucleotides intended for use as antisense oligonucleotides may be modified to increase stability and half-life. Possible modifications include, but are not limited to phosphorothioate or 2′ O-methyl linkages, and the inclusion of nontraditional bases such as inosine and queosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytosine, guanine, thymine, and uracil which are not as easily recognized by endogenous nucleases.

The invention also provides an isolated polypeptide comprising a polymorphic variant of (a) the reference CYP1B1 amino acid sequence shown in Figure 3 or (b) a fragment of this reference sequence. The location of a variant amino acid in a CYP1B1 polypeptide or fragment of the invention is preferably identified by aligning its sequence against SEQ ID NO:3 (Fig. 3). A CYP1B1 protein variant of the invention comprises an amino acid sequence identical to SEQ ID NO:3 for those regions of SEQ ID NO:3 that are encoded by examined portions of the CYP1B1 gene (as described in the Examples below), except for having glycine at a position corresponding to amino acid position 443, and may also comprise one or more additional variant amino acids selected from the group consisting of glycine at a position corresponding to amino acid position 48, serine at a position corresponding to amino acid position 119, leucine at a position corresponding to amino acid position 432 and serine at a position corresponding to amino acid position 453. Thus, a CYP1B1 fragment of the invention, also referred to herein as a CYP1B1 peptide variant, is any fragment of a CYP1B1 protein variant that contains glycine at a position corresponding to amino acid position 443. The invention specifically excludes amino acid sequences identical to those previously identified for CYP1B1, including SEQ ID NO:3, and previously described fragments thereof. CYP1B1 protein variants included within the invention comprise all amino acid sequences based on SEQ ID NO:3 and having the combination of amino acid variations described in Table 2 below. In preferred embodiments, a CYP1B1 protein variant of the invention is encoded by an isogene defined by one of the observed haplotypes, 1, 3-11 and 13-20, shown in Table 5.

Table 2. Novel Polymorphic Variants of CYP1B1

Polymorphic Variant Number	Amino Acid Position and Identities				
	48	119	432	443	453
1	R	A	V	G	N
2	R	A	V	G	S
3	R	A	L	G	N
4	R	A	L	G	S
5	R	S	V	G	N
6	R	S	V	G	S
7	R	S	L	G	N
8	R	S	L	G	S
9	G	A	V	G	N
10	G	A	V	G	S
11	G	A	L	G	N
12	G	A	L	G	S
13	G	S	V	G	N
14	G	S	V	G	S
15	G	S	L	G	N
16	G	S	L	G	S

A CYP1B1 peptide variant of the invention is at least 6 amino acids in length and is preferably any number between 6 and 30 amino acids long, more preferably between 10 and 25, and most preferably between 15 and 20 amino acids long. Such CYP1B1 peptide variants may be useful as antigens to generate antibodies specific for one of the above CYP1B1 isoforms. In addition, the CYP1B1 peptide variants may be useful in drug screening assays.

A CYP1B1 variant protein or peptide of the invention may be prepared by chemical synthesis or by expressing an appropriate variant CYP1B1 genomic or cDNA sequence described above. Alternatively, the CYP1B1 protein variant may be isolated from a biological sample of an individual having a CYP1B1 isogene which encodes the variant protein. Where the sample contains two different CYP1B1 isoforms (i.e., the individual has different CYP1B1 isogenes), a particular CYP1B1 isoform of the invention can be isolated by immunoaffinity chromatography using an antibody which specifically binds to that particular CYP1B1 isoform but does not bind to the other CYP1B1 isoform.

The expressed or isolated CYP1B1 protein or peptide may be detected by methods known in the art, including Coomassie blue staining, silver staining, and Western blot analysis using antibodies specific for the isoform of the CYP1B1 protein or peptide as discussed further below. CYP1B1 variant proteins and peptides can be purified by standard protein purification procedures known in the art, including differential precipitation, molecular sieve chromatography, ion-exchange chromatography, isoelectric focusing, gel electrophoresis, affinity and immunoaffinity chromatography and the like. (Ausubel et. al., 1987, In Current Protocols in Molecular Biology John Wiley and Sons, New York, New York). In the case of immunoaffinity chromatography, antibodies specific for a particular polymorphic variant may be used.

A polymorphic variant CYP1B1 gene of the invention may also be fused in frame with a

heterologous sequence to encode a chimeric CYP1B1 protein. The non-CYP1B1 portion of the chimeric protein may be recognized by a commercially available antibody. In addition, the chimeric protein may also be engineered to contain a cleavage site located between the CYP1B1 and non-CYP1B1 portions so that the CYP1B1 protein may be cleaved and purified away from the non-CYP1B1 portion.

An additional embodiment of the invention relates to using a novel CYP1B1 protein isoform, or a fragment thereof, in any of a variety of drug screening assays. Such screening assays may be performed to identify agents that bind specifically to all known CYP1B1 protein isoforms or to only a subset of one or more of these isoforms. The agents may be from chemical compound libraries, peptide libraries and the like. The CYP1B1 protein or peptide variant may be free in solution or affixed to a solid support. In one embodiment, high throughput screening of compounds for binding to a CYP1B1 variant may be accomplished using the method described in PCT application WO84/03565, in which large numbers of test compounds are synthesized on a solid substrate, such as plastic pins or some other surface, contacted with the CYP1B1 protein(s) of interest and then washed. Bound CYP1B1 protein(s) are then detected using methods well-known in the art.

In another embodiment, a novel CYP1B1 protein isoform may be used in assays to measure the binding affinities of one or more candidate drugs targeting the CYP1B1 protein or to measure the enzymatic activity of CYP1B1 when using one or more candidate drugs as substrates.

In yet another embodiment, when a particular CYP1B1 haplotype or group of CYP1B1 haplotypes encodes a CYP1B1 protein variant with an amino acid sequence distinct from that of CYP1B1 protein isoforms encoded by other CYP1B1 haplotypes, then detection of that particular CYP1B1 haplotype or group of CYP1B1 haplotypes may be accomplished by detecting expression of the encoded CYP1B1 protein variant using any of the methods described herein or otherwise commonly known to the skilled artisan.

In another embodiment, the invention provides antibodies specific for and immunoreactive with one or more of the novel CYP1B1 protein or peptide variants described herein. The antibodies may be either monoclonal or polyclonal in origin. The CYP1B1 protein or peptide variant used to generate the antibodies may be from natural or recombinant sources (in vitro or in vivo) or produced by chemical synthesis or semi-synthetic synthesis using synthesis techniques known in the art. If the CYP1B1 protein or peptide variant is of insufficient size to be antigenic, it may be concatenated or conjugated, complexed, or otherwise covalently linked to a carrier molecule to enhance the antigenicity of the peptide. Examples of carrier molecules, include, but are not limited to, albumins (e.g., human, bovine, fish, ovine), and keyhole limpet hemocyanin (Basic and Clinical Immunology, 1991, Eds. D.P. Stites, and A.I. Terr, Appleton and Lange, Norwalk Connecticut, San Mateo, California).

In one embodiment, an antibody specifically immunoreactive with one of the novel protein or peptide variants described herein is administered to an individual to neutralize activity of the CYP1B1

isoform expressed by that individual. The antibody may be formulated as a pharmaceutical composition which includes a pharmaceutically acceptable carrier.

Antibodies specific for and immunoreactive with one of the novel protein isoforms described herein may be used to immunoprecipitate the CYP1B1 protein variant from solution as well as react with CYP1B1 protein isoforms on Western or immunoblots of polyacrylamide gels on membrane supports or substrates. In another preferred embodiment, the antibodies will detect CYP1B1 protein isoforms in paraffin or frozen tissue sections, or in cells which have been fixed or unfixed and prepared on slides, coverslips, or the like, for use in immunocytochemical, immunohistochemical, and immunofluorescence techniques.

In another embodiment, an antibody specifically immunoreactive with one of the novel CYP1B1 protein variants described herein is used in immunoassays to detect this variant in biological samples. In this method, an antibody of the present invention is contacted with a biological sample and the formation of a complex between the CYP1B1 protein variant and the antibody is detected. As described, suitable immunoassays include radioimmunoassay, Western blot assay, immunofluorescent assay, enzyme linked immunoassay (ELISA), chemiluminescent assay, immunohistochemical assay, immunocytochemical assay, and the like (see, e.g., Principles and Practice of Immunoassay, 1991, Eds. Christopher P. Price and David J. Neoman, Stockton Press, New York, New York; Current Protocols in Molecular Biology, 1987, Eds. Ausubel et al., John Wiley and Sons, New York, New York). Standard techniques known in the art for ELISA are described in Methods in Immunodiagnosis, 2nd Ed., Eds. Rose and Bigazzi, John Wiley and Sons, New York 1980; and Campbell et al., 1984, Methods in Immunology, W.A. Benjamin, Inc.). Such assays may be direct, indirect, competitive, or noncompetitive as described in the art (see, e.g., Principles and Practice of Immunoassay, 1991, Eds. Christopher P. Price and David J. Neoman, Stockton Pres, NY, NY; and Oellirich, M., 1984, J. Clin. Chem. Clin. Biochem., 22:895-904). Proteins may be isolated from test specimens and biological samples by conventional methods, as described in Current Protocols in Molecular Biology, supra.

Exemplary antibody molecules for use in the detection and therapy methods of the present invention are intact immunoglobulin molecules, substantially intact immunoglobulin molecules, or those portions of immunoglobulin molecules that contain the antigen binding site. Polyclonal or monoclonal antibodies may be produced by methods conventionally known in the art (e.g., Kohler and Milstein, 1975, Nature, 256:495-497; Campbell Monoclonal Antibody Technology, the Production and Characterization of Rodent and Human Hybridomas, 1985, In: Laboratory Techniques in Biochemistry and Molecular Biology, Eds. Burdon et al., Volume 13, Elsevier Science Publishers, Amsterdam). The antibodies or antigen binding fragments thereof may also be produced by genetic engineering. The technology for expression of both heavy and light chain genes in E. coli is the subject of PCT patent applications, publication number WO 901443, WO 901443 and WO 9014424 and in Huse et al., 1989, Science, 246:1275-1281. The antibodies may also be humanized (e.g.,

Queen, C. et al. 1989 Proc. Natl. Acad. Sci.USA 86;10029).

Effect(s) of the polymorphisms identified herein on expression of CYP1B1 may be investigated by various means known in the art, such as by *in vitro* translation of mRNA transcripts of the CYP1B1 gene, cDNA orfragment thereof, or by preparing recombinant cells and/or nonhuman recombinant organisms, preferably recombinant animals, containing a polymorphic variant of the CYP1B1 gene. As used herein, "expression" includes but is not limited to one or more of the following: transcription of the gene into precursor mRNA; splicing and other processing of the precursor mRNA to produce mature mRNA; mRNA stability; translation of the mature mRNA(s) into CYP1B1 protein(s) (including effects of polymorphihsms on codon usage and tRNA availability); and glycosylation and/or other modifications of the translation product, if required for proper expression and function.

To prepare a recombinant cell of the invention, the desired CYP1B1 isogene, cDNA or coding sequence may be introduced into the cell in a vector such that the isogene, cDNA or coding sequence remains extrachromosomal. In such a situation, the gene will be expressed by the cell from the extrachromosomal location. In a preferred embodiment, the CYP1B1 isogene, cDNA or coding sequence is introduced into a cell in such a way that it recombines with the endogenous CYP1B1 gene present in the cell. Such recombination requires the occurrence of a double recombination event, thereby resulting in the desired CYP1B1 gene polymorphism. Vectors for the introduction of genes both for recombination and for extrachromosomal maintenance are known in the art, and any suitable vector or vector construct may be used in the invention. Methods such as electroporation, particle bombardment, calcium phosphate co-precipitation and viral transduction for introducing DNA into cells are known in the art; therefore, the choice of method may lie with the competence and preference of the skilled practitioner. Examples of cells into which the CYP1B1 isogene, cDNA or coding sequence may be introduced include, but are not limited to, continuous culture cells, such as COS, CHO, NIH/3T3, and primary or culture cells of the relevant tissue type, i.e., they express the CYP1B1 isogene, cDNA or coding sequence. Such recombinant cells can be used to compare the biological activities of the different protein variants.

Recombinant nonhuman organisms, i.e., transgenic animals, expressing a variant CYP1B1 gene, cDNA or coding sequence are prepared using standard procedures known in the art. Preferably, a construct comprising the variant gene, cDNA or coding sequence is introduced into a nonhuman animal or an ancestor of the animal at an embryonic stage, i.e., the one-cell stage, or generally not later than about the eight-cell stage. Transgenic animals carrying the constructs of the invention can be made by several methods known to those having skill in the art. One method involves transfecting into the embryo a retrovirus constructed to contain one or more insulator elements, a gene or genes (or cDNA or coding sequence) of interest, and other components known to those skilled in the art to provide a complete shuttle vector harboring the insulated gene(s) as a transgene, see e.g., U.S. Patent No. 5,610,053. Another method involves directly injecting a transgene into the embryo. A third

method involves the use of embryonic stem cells. Examples of animals into which the CYP1B1 isogene, cDNA or coding sequences may be introduced include, but are not limited to, mice, rats, other rodents, and nonhuman primates (see "The Introduction of Foreign Genes into Mice" and the cited references therein, In: Recombinant DNA, Eds. J.D. Watson, M. Gilman, J. Witkowski, and M. Zoller; W.H. Freeman and Company, New York, pages 254-272). Transgenic animals stably expressing a human CYP1B1 isogene, cDNA or coding sequence and producing the encoded human CYP1B1 protein can be used as biological models for studying diseases related to abnormal CYP1B1 expression and/or activity, and for screening and assaying various candidate drugs, compounds, and treatment regimens to reduce the symptoms or effects of these diseases.

An additional embodiment of the invention relates to pharmaceutical compositions for treating disorders affected by expression or function of a novel CYP1B1 isogene described herein. The pharmaceutical composition may comprise any of the following active ingredients: a polynucleotide comprising one of these novel CYP1B1 isogenes (or cDNAs or coding sequences); an antisense oligonucleotide directed against one of the novel CYP1B1 isogenes, a polynucleotide encoding such an antisense oligonucleotide, or another compound which inhibits expression of a novel CYP1B1 isogene described herein. Preferably, the composition contains the active ingredient in a therapeutically effective amount. By therapeutically effective amount is meant that one or more of the symptoms relating to disorders affected by expression or function of a novel CYP1B1 isogene is reduced and/or eliminated. The composition also comprises a pharmaceutically acceptable carrier, examples of which include, but are not limited to, saline, buffered saline, dextrose, and water. Those skilled in the art may employ a formulation most suitable for the active ingredient, whether it is a polynucleotide, oligonucleotide, protein, peptide or small molecule antagonist. The pharmaceutical composition may be administered alone or in combination with at least one other agent, such as a stabilizing compound. Administration of the pharmaceutical composition may be by any number of routes including, but not limited to oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, intradermal, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing Co., Easton, PA).

For any composition, determination of the therapeutically effective dose of active ingredient and/or the appropriate route of administration is well within the capability of those skilled in the art. For example, the dose can be estimated initially either in cell culture assays or in animal models. The animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans. The exact dosage will be determined by the practitioner, in light of factors relating to the patient requiring treatment, including but not limited to severity of the disease state, general health, age, weight and gender of the patient, diet, time and frequency of administration, other

drugs being taken by the patient, and tolerance/response to the treatment.

Any or all analytical and mathematical operations involved in practicing the methods of the present invention may be implemented by a computer. In addition, the computer may execute a program that generates views (or screens) displayed on a display device and with which the user can interact to view and analyze large amounts of information relating to the CYP1B1 gene and its genomic variation, including chromosome location, gene structure, and gene family, gene expression data, polymorphism data, genetic sequence data, and clinical data population data (e.g., data on ethnogeographic origin, clinical responses, genotypes, and haplotypes for one or more populations). The CYP1B1 polymorphism data described herein may be stored as part of a relational database (e.g., an instance of an Oracle database or a set of ASCII flat files). These polymorphism data may be stored on the computer's hard drive or may, for example, be stored on a CD-ROM or on one or more other storage devices accessible by the computer. For example, the data may be stored on one or more databases in communication with the computer via a network.

Preferred embodiments of the invention are described in the following examples. Other embodiments within the scope of the claims herein will be apparent to one skilled in the art from consideration of the specification or practice of the invention as disclosed herein. It is intended that the specification, together with the examples, be considered exemplary only, with the scope and spirit of the invention being indicated by the claims which follow the examples.

EXAMPLES

The Examples herein are meant to exemplify the various aspects of carrying out the invention and are not intended to limit the scope of the invention in any way. The Examples do not include detailed descriptions for conventional methods employed, such as in the performance of genomic DNA isolation, PCR and sequencing procedures. Such methods are well-known to those skilled in the art and are described in numerous publications, for example, Sambrook, Fritsch, and Maniatis, "Molecular Cloning: A Laboratory Manual", 2nd Edition, Cold Spring Harbor Laboratory Press, USA, (1989).

EXAMPLE 1

This example illustrates examination of various regions of the CYP1B1 gene for polymorphic sites.

Amplification of Target Regions

The following target regions of the CYP1B1 gene were amplified using PCR primer pairs. The primers used for each region are represented below by providing the nucleotide positions of their initial and final nucleotides, which correspond to positions in SEQ ID NO:1 (Figure 1).

PCR Primer Pairs

Fragment No.	Forward Primer	Reverse Primer	PCR Product
Fragment 1	882-903	complement of 1559-1540	678 nt
Fragment 2	1319-1340	complement of 1955-1932	637 nt
Fragment 3	2284-2306	complement of 2884-2865	601 nt
Fragment 4	2526-2547	complement of 3165-3146	640 nt
Fragment 5	2830-2849	complement of 3446-3425	617 nt
Fragment 6	3080-3099	complement of 3779-3757	700 nt
Fragment 7	6304-6329	complement of 7013-6993	710 nt
Fragment 8	6680-6703	complement of 7308-7287	629 nt

These primer pairs were used in PCR reactions containing genomic DNA isolated from immortalized cell lines for each member of the Index Repository. The PCR reactions were carried out under the following conditions:

Reaction volume	= 10 µl
10 x Advantage 2 Polymerase reaction buffer (Clontech)	= 1 µl
100 ng of human genomic DNA	= 1 µl
10 mM dNTP	= 0.4 µl
Advantage 2 Polymerase enzyme mix (Clontech)	= 0.2 µl
Forward Primer (10 µM)	= 0.4 µl
Reverse Primer (10 µM)	= 0.4 µl
Water	= 6.6µl

Amplification profile:

97°C - 2 min. 1 cycle

97°C - 15 sec.
70°C - 45 sec.
72°C - 45 sec. } 10 cycles

97°C - 15 sec.
64°C - 45 sec.
72°C - 45 sec. } 35 cycles

Sequencing of PCR Products

The PCR products were purified using a Whatman/Polyfiltronics 100 µl 384 well unifilter plate essentially according to the manufacturers protocol. The purified DNA was eluted in 50 µl of distilled water. Sequencing reactions were set up using Applied Biosystems Big Dye Terminator chemistry essentially according to the manufacturers protocol. The purified PCR products were sequenced in both directions using the primer sets described previously or those represented below by the nucleotide positions of their initial and final nucleotides, which correspond to positions in SEQ ID NO:1 (Figure 1). Reaction products were purified by isopropanol precipitation, and run on an Applied Biosystems 3700 DNA Analyzer.

Sequencing Primer Pairs

Fragment No.	Forward Primer	Reverse Primer
Fragment 1	1021-1039	complement of 1521-1502
Fragment 2	1361-1380	complement of 1896-1878
Fragment 3	2330-2350	complement of 2843-2825
Fragment 4	2554-2572	complement of 3100-3081
Fragment 5	2878-2897	complement of 3359-3338
Fragment 6	3136-3155	complement of 3641-3622
Fragment 7	6481-6501	complement of 6942-6924
Fragment 8	6710-6729	complement of 7229-7208

Analysis of Sequences for Polymorphic Sites

Sequence information for a minimum of 80 humans was analyzed for the presence of polymorphisms using the Polyphred program (Nickerson et al., *Nucleic Acids Res.* 14:2745-2751, 1997). The presence of a polymorphism was confirmed on both strands. The polymorphisms and their locations in the CYP1B1 reference genomic sequence (SEQ ID NO:1) are listed in Table 3 below.

Table 3. Polymorphic Sites Identified in the CYP1B1 Gene

Polymorphic Site Number	PolyId(a)	Nucleotide Position	Reference Allele	Variant Allele	CDS Variant Position	AA Variant
PS1	1834605	1063	C	T		
PS2	1834603	1134	T	C		
PS3	1834599	1342	G	A		
PS4	1834597	1357	T	C		
PS5	1834593	1468	C	T		
PS6	1834589	2454	C	T		
PS7	1834587	2456	C	T		
PS8(R)	1834585	2610	C	G	142	R48G
PS9(R)	1834579	2823	G	T	355	A119S
PS10	1834577	3032	C	A	564	G188G
PS11	1834575	3197	G	C	729	V243V
PS12	1834569	3551	C	T		
PS13	1834565	6551	T	C	1047	Y349Y
PS14	1834563	6665	A	G	1161	E387E
PS15(R)	1834561	6798	G	C	1294	V432L
PS16	1834559	6832	C	G	1328	A443G
PS17(R)	1834557	6851	T	C	1347	D449D
PS18(R)	1834553	6862	A	G	1358	N453S
PS19	1834551	7242	G	A		
PS20(R)	1834549	7254	C	G		

(a) PolyId is a unique identifier assigned to each PS by Genaissance Pharmaceuticals, Inc.
(R) Reported previously.

EXAMPLE 2

This example illustrates analysis of the CYP1B1 polymorphisms identified in the Index Repository for human genotypes and haplotypes.

The different genotypes containing these polymorphisms that were observed in unrelated members of the reference population are shown in Table 4 below, with the haplotype pair indicating the combination of haplotypes determined for the individual using the haplotype derivation protocol described below. In Table 4, homozygous positions are indicated by one nucleotide and heterozygous positions are indicated by two nucleotides. Missing nucleotides in any given genotype in Table 4 were inferred based on linkage disequilibrium and/or Mendelian inheritance.

Table 4 (Part 1). Genotypes and Haplotype Pairs Observed for CYP1B1 Gene

Genotype Number	HAP Pair		Polymorphic Sites									
			PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	PS10
1	5	5	C	T	G	C	T	C	T	G	T	C
2	11	11	C	T	G	T	C	C	C	C	G	C
3	10	10	C	T	G	T	C	C	C	C	G	C
4	12	12	C	T	G	T	C	C	C	C	G	C
5	10	11	C	T	G	T	C	C	C	C	G	C
6	11	4	C	T	G	T/C	C/T	C	C	C	G	C
7	12	18	C	T	G	T	C	C	C/T	C/G	G/T	C
8	12	20	C/T	T	G	T	C	C/T	C	C/G	G/T	C
9	10	14	C	T	G	T	C	C	C	C/G	G/T	C/A
10	10	8	C	T	G	T	C	C	C	C	G	C
11	10	3	C	T	G	T/C	C	C	C	C/G	G	C
12	10	1	C	T/C	G	T	C	C	C	C	G	C
13	11	7	C	T	G	T/C	C/T	C	C/T	C/G	G/T	C
14	20	13	T/C	T	G	T	C	T/C	C	G	T/G	C
15	5	20	C/T	T	G	C/T	T/C	C/T	T/C	G	T	C
16	11	20	C/T	T	G	T	C	C/T	C	C/G	G/T	C
17	12	11	C	T	G	T	C	C	C	C	G	C
18	10	17	C	T	G	T	C	C	C/T	C/G	G/T	C
19	20	16	T/C	T	G	T	C	T/C	C	G	T	C
20	17	14	C	T	G	T	C	C	T/C	G	T	C/A
21	11	13	C	T	G	T	C	C	C	C/G	G	C
22	11	14	C	T	G	T	C	C	C	C/G	G/T	C/A
23	12	17	C	T	G	T	C	C	C/T	C/G	G/T	C
24	10	12	C	T	G	T	C	C	C	C	G	C
25	17	13	C	T	G	T	C	C	T/C	G	T/G	C
26	10	2	C	T	G/A	T	C	C	C	C	G	C
27	10	9	C	T	G	T	C	C	C	C	G	C
28	5	11	C	T	G	C/T	T/C	C	T/C	G/C	T/G	C
29	12	5	C	T	G	T/C	C/T	C	C/T	C/G	G/T	C
30	5	6	C	T	G	C	T	C	T	G	T	C
31	10	5	C	T	G	T/C	C/T	C	C/T	C/G	G/T	C
32	19	15	C	T	G	T	C	C	T/C	G	T	C
33	12	14	C	T	G	T	C	C	C	C/G	G/T	C/A

Table 4 (Part 2). Genotypes and Haplotype Pairs Observed for CYP1B1 Gene

Genotype Number	HAP Pair		Polymorphic Sites PS11	PS12	PS13	PS14	PS15	PS16	PS17	PS18	PS19	PS20
1	5	5	G	C	T	A	C	C	C	A	G	C
2	11	11	G	C	T	A	C	C	C	G	G	C
3	10	10	G	C	T	A	C	C	C	A	G	C
4	12	12	G	C	T	A	G	C	T	A	G	C
5	10	11	G	C	T	A	C	C	C	A/G	G	C
6	11	4	G	C	T	A	C	C	C	G	G	C
7	12	18	G	C	T	A	G	C	T	A	G	C
8	12	20	G	C	T	A	G/C	C	T/C	A	G	C
9	10	14	G	C	T	A	C/G	C	C/T	A	G	C
10	10	8	G	C	T/C	A	C	C	C	A	G	C
11	10	3	G	C	T	A	C/G	C	C/T	A	G	C
12	10	1	G	C	T	A	C	C	C	A	G	C
13	11	7	G	C	T	A/G	C	C	C	G/A	G	C
14	20	13	G/C	C	T	A	C/G	C	C/T	A	G	C
15	5	20	G	C	T	A	C	C	C	A	G	C
16	11	20	G	C	T	A	C	C	C	G/A	G	C
17	12	11	G	C	T	A	G/C	C	T/C	A/G	G	C
18	10	17	G/C	C	T	A	C/G	C	C/T	A	G	C
19	20	16	G	C	T	A	C/G	C	C/T	A	G	C
20	17	14	C/G	C	T	A	G	C	T	A	G	C
21	11	13	G/C	C	T	A	C/G	C	C/T	G/A	G	C
22	11	14	G	C	T	A	C/G	C	C/T	G/A	G	C
23	12	17	G/C	C	T	A	G	C	T	A	G	C
24	10	12	G	C	T	A	C/G	C	C/T	A	G	C
25	17	13	C	C	T	A	G	C	T	A	G	C
26	10	2	G	C	T	A	C/G	C	C/T	A	G	C
27	10	9	G	C	T	A	C	C	C	A	G/A	C
28	5	11	G	C	T	A	C	C	C	A/G	G	C
29	12	5	G	C	T	A	G/C	C	T/C	A	G	C
30	5	6	G	C	T	A	C	C	C	A	G	C/G
31	10	5	G	C	T	A	C	C	C	A	G	C
32	19	15	G/C	T/C	T	A	C/G	C/G	C/T	A	G	C
33	12	14	G	C	T	A	G	C	T	A	G	C

The haplotype pairs shown in Table 4 were estimated from the unphased genotypes using a computer-implemented extension of Clark's algorithm (Clark, A.G. 1990 *Mol Bio Evol* 7, 111-122) for assigning haplotypes to unrelated individuals in a population sample, as described in PCT/US01/12831, filed April 18, 2001. In this method, haplotypes are assigned directly from individuals who are homozygous at all sites or heterozygous at no more than one of the variable sites. This list of haplotypes is then used to deconvolute the unphased genotypes in the remaining (multiply heterozygous) individuals. In the present analysis, the list of haplotypes was augmented with haplotypes obtained from two families (one three-generation Caucasian family and one two-generation African-American family).

By following this protocol, it was determined that the Index Repository examined herein and, by extension, the general population contains the 20 human CYP1B1 haplotypes shown in Table 5

below.

A CYP1B1 isogene defined by a full-haplotype shown in Table 5 below comprises the regions of the SEQ ID NOS indicated in Table 5, with their corresponding set of polymorphic locations and identities, which are also set forth in Table 5.

Table 5 (Part 1). Haplotypes of the CYP1B1 gene.

Regions Examined(a)	PS No.(b)	PS Position(c)	Haplotype Number(d) 1	2	3	4	5	6	7	8	9	10
882-1955	1	1063/30	C	C	C	C	C	C	C	C	C	C
882-1955	2	1134/150	C	T	T	T	T	T	T	T	T	T
882-1955	3	1342/270	G	A	G	G	G	G	G	G	G	G
882-1955	4	1357/390	T	T	C	C	C	C	C	T	T	T
882-1955	5	1468/510	C	C	C	T	T	T	T	C	C	C
2284-3779	6	2454/630	C	C	C	C	C	C	C	C	C	C
2284-3779	7	2456/750	C	C	C	C	T	T	T	C	C	C
2284-3779	8	2610/870	C	C	G	C	G	G	G	C	C	C
2284-3779	9	2823/990	G	G	G	G	T	T	T	G	G	G
2284-3779	10	3032/1110	C	C	C	C	C	C	C	C	C	C
2284-3779	11	3197/1230	G	G	G	G	G	G	G	G	G	G
2284-3779	12	3551/1350	C	C	C	C	C	C	C	C	C	C
6304-7308	13	6551/1470	T	T	T	T	T	T	T	C	T	T
6304-7308	14	6665/1590	A	A	A	A	A	A	G	A	A	A
6304-7308	15	6798/1710	C	G	G	C	C	C	C	C	C	C
6304-7308	16	6832/1830	C	C	C	C	C	C	C	C	C	C
6304-7308	17	6851/1950	C	T	T	C	C	C	C	C	C	C
6304-7308	18	6862/2070	A	A	A	G	A	A	A	A	A	A
6304-7308	19	7242/2190	G	G	G	G	G	G	G	G	A	G
6304-7308	20	7254/2310	C	C	C	C	C	G	C	C	C	C

(a) Region examined represents the nucleotide positions defining the start and stop positions within SEQ ID NO:1 of the regions sequenced;
(b) PS = polymorphic site;
(c) Position of PS within the indicated SEQ ID NO, with the 1st position number referring to SEQ ID NO:1 and the 2nd position number referring to SEQ ID NO:74, a modified version of SEQ ID NO:1 that comprises the context sequence of each polymorphic site, PS1-PS20, to facilitate electronic searching of the haplotypes;
(d) Alleles for CYP1B1 haplotypes are presented 5′ to 3′ in each column.

Table 5 (Part 2). Haplotypes of the CYP1B1 gene.

Regions Examined(a)	PS No.(b)	PS Position(c)	Haplotype Number(d) 11	12	13	14	15	16	17	18	19	20
882-1955	1	1063/30	C	C	C	C	C	C	C	C	C	T
882-1955	2	1134/150	T	T	T	T	T	T	T	T	T	T
882-1955	3	1342/270	G	G	G	G	G	G	G	G	G	G
882-1955	4	1357/390	T	T	T	T	T	T	T	T	T	T
882-1955	5	1468/510	C	C	C	C	C	C	C	C	C	C
2284-3779	6	2454/630	C	C	C	C	C	C	C	C	C	T
2284-3779	7	2456/750	C	C	C	C	C	C	T	T	T	C
2284-3779	8	2610/870	C	C	G	G	G	G	G	G	G	G
2284-3779	9	2823/990	G	G	G	T	T	T	T	T	T	T
2284-3779	10	3032/1110	C	C	C	A	C	C	C	C	C	C
2284-3779	11	3197/1230	G	G	C	G	C	G	C	G	G	G
2284-3779	12	3551/1350	C	C	C	C	C	C	C	C	T	C
6304-7308	13	6551/1470	T	T	T	T	T	T	T	T	T	T
6304-7308	14	6665/1590	A	A	A	A	A	A	A	A	A	A
6304-7308	15	6798/1710	C	G	G	G	G	G	G	G	C	C
6304-7308	16	6832/1830	C	C	C	C	G	C	C	C	C	C
6304-7308	17	6851/1950	C	T	T	T	T	T	T	T	C	C
6304-7308	18	6862/2070	G	A	A	A	A	A	A	A	A	A
6304-7308	19	7242/2190	G	G	G	G	G	G	G	G	G	G
6304-7308	20	7254/2310	C	C	C	C	C	C	C	C	C	C

(a) Region examined represents the nucleotide positions defining the start and stop positions within SEQ ID NO:1 of the regions sequenced;
(b) PS = polymorphic site;
(c) Position of PS within the indicated SEQ ID NO, with the 1st position number referring to SEQ ID NO:1 and the 2nd position number referring to SEQ ID NO:74, a modified version of SEQ ID NO:1 that comprises the context sequence of each polymorphic site, PS1-PS20, to facilitate electronic searching of the haplotypes;
(d) Alleles for CYP1B1 haplotypes are presented 5′ to 3′ in each column.

SEQ ID NO:1 refers to Figure 1, with the two alternative allelic variants of each polymorphic site indicated by the appropriate nucleotide symbol. SEQ ID NO:74 is a modified version of SEQ ID NO:1 that shows the context sequence of each of PS1-PS20 in a uniform format to facilitate electronic searching of the CYP1B1 haplotypes. For each polymorphic site, SEQ ID NO:74 contains a block of 60 bases of the nucleotide sequence encompassing the centrally-located polymorphic site at the 30th position, followed by 60 bases of unspecified sequence to represent that each polymorphic site is separated by genomic sequence whose composition is defined elsewhere herein.

Table 6 below shows the percent of chromosomes characterized by a given CYP1B1 haplotype for all unrelated individuals in the Index Repository for which haplotype data was obtained. The percent of these unrelated individuals who have a given CYP1B1 haplotype pair is shown in Table 7. In Tables 6 and 7, the "Total" column shows this frequency data for all of these unrelated individuals, while the other columns show the frequency data for these unrelated individuals categorized according to their self-identified ethnogeographic origin. Abbreviations used in Tables 6 and 7 are AF = African Descent, AS = Asian, CA = Caucasian, HL = Hispanic-Latino, and AM =

Native American.

Table 6. Frequency of Observed CYP1B1 Haplotypes In Unrelated Individuals

HAP No.	HAP ID	Total	CA	AF	AS	HL	AM
1	1837328	0.61	0.0	0.0	2.5	0.0	0.0
2	1837331	0.61	2.38	0.0	0.0	0.0	0.0
3	1837332	0.61	0.0	0.0	0.0	2.78	0.0
4	1837333	0.61	0.0	0.0	0.0	2.78	0.0
5	1837316	19.51	23.81	7.5	17.5	27.78	33.33
6	1837326	0.61	0.0	0.0	0.0	2.78	0.0
7	1837324	0.61	2.38	0.0	0.0	0.0	0.0
8	1837327	0.61	0.0	0.0	2.5	0.0	0.0
9	1837329	0.61	0.0	0.0	2.5	0.0	0.0
10	1837314	27.44	9.52	7.5	72.5	19.44	33.33
11	1837317	10.98	23.81	5.0	0.0	11.11	33.33
12	1837315	21.95	35.71	25.0	2.5	27.78	0.0
13	1837321	2.44	0.0	10.0	0.0	0.0	0.0
14	1837319	3.66	0.0	12.5	0.0	2.78	0.0
15	1837325	0.61	0.0	2.5	0.0	0.0	0.0
16	1837330	0.61	0.0	2.5	0.0	0.0	0.0
17	1837318	3.66	0.0	15.0	0.0	0.0	0.0
18	1837323	0.61	0.0	2.5	0.0	0.0	0.0
19	1837322	0.61	0.0	2.5	0.0	0.0	0.0
20	1837320	3.05	2.38	7.5	0.0	2.78	0.0

Table 7. Frequency of Observed CYP1B1 Haplotype Pairs In Unrelated Individuals

HAP1	HAP2	Total	CA	AF	AS	HL	AM
5	5	4.88	0.0	5.0	10.0	5.56	0.0
11	11	2.44	4.76	0.0	0.0	0.0	33.33
10	10	13.41	0.0	0.0	55.0	0.0	0.0
12	12	6.1	9.52	5.0	0.0	11.11	0.0
10	11	2.44	0.0	0.0	0.0	11.11	0.0
11	4	1.22	0.0	0.0	0.0	5.56	0.0
12	18	1.22	0.0	5.0	0.0	0.0	0.0
12	20	1.22	0.0	5.0	0.0	0.0	0.0
10	14	2.44	0.0	5.0	0.0	5.56	0.0
10	8	1.22	0.0	0.0	5.0	0.0	0.0
10	3	1.22	0.0	0.0	0.0	5.56	0.0
10	1	1.22	0.0	0.0	5.0	0.0	0.0
11	7	1.22	4.76	0.0	0.0	0.0	0.0
20	13	1.22	0.0	5.0	0.0	0.0	0.0
5	20	1.22	0.0	0.0	0.0	5.56	0.0
11	20	1.22	4.76	0.0	0.0	0.0	0.0
12	11	6.1	19.05	0.0	0.0	5.56	0.0
10	17	1.22	0.0	5.0	0.0	0.0	0.0
20	16	1.22	0.0	5.0	0.0	0.0	0.0
17	14	1.22	0.0	5.0	0.0	0.0	0.0
11	13	1.22	0.0	5.0	0.0	0.0	0.0
11	14	1.22	0.0	5.0	0.0	0.0	0.0
12	17	2.44	0.0	10.0	0.0	0.0	0.0
10	12	4.88	4.76	5.0	5.0	5.56	0.0
17	13	2.44	0.0	10.0	0.0	0.0	0.0
10	2	1.22	4.76	0.0	0.0	0.0	0.0
10	9	1.22	0.0	0.0	5.0	0.0	0.0
5	11	2.44	9.52	0.0	0.0	0.0	0.0
12	5	13.41	28.57	5.0	0.0	22.22	0.0
5	6	1.22	0.0	0.0	0.0	5.56	0.0
10	5	10.98	9.52	0.0	15.0	11.11	66.67
19	15	1.22	0.0	5.0	0.0	0.0	0.0
12	14	2.44	0.0	10.0	0.0	0.0	0.0

The size and composition of the Index Repository were chosen to represent the genetic diversity across and within four major population groups comprising the general United States population. For example, as described in Table 1 above, this repository contains approximately equal sample sizes of African-descent, Asian-American, European-American, and Hispanic-Latino population groups. Almost all individuals representing each group had all four grandparents with the same ethnogeographic background. The number of unrelated individuals in the Index Repository provides a sample size that is sufficient to detect SNPs and haplotypes that occur in the general population with high statistical certainty. For instance, a haplotype that occurs with a frequency of 5% in the general population has a probability higher than 99.9% of being observed in a sample of 80 individuals from the general population. Similarly, a haplotype that occurs with a frequency of 10% in a specific population group has a 99% probability of being observed in a sample of 20 individuals from that population group. In addition, the size and composition of the Index Repository means that

the relative frequencies determined therein for the haplotypes and haplotype pairs of the CYP1B1 gene are likely to be similar to the relative frequencies of these CYP1B1 haplotypes and haplotype pairs in the general U.S. population and in the four population groups represented in the Index Repository. The genetic diversity observed for the three Native Americans is presented because it is of scientific interest, but due to the small sample size it lacks statistical significance.

In view of the above, it will be seen that the several advantages of the invention are achieved and other advantageous results attained.

As various changes could be made in the above methods and compositions without departing from the scope of the invention, it is intended that all matter contained in the above description and shown in the accompanying drawings shall be interpreted as illustrative and not in a limiting sense.

All references cited in this specification, including patents and patent applications, are hereby incorporated in their entirety by reference. The discussion of references herein is intended merely to summarize the assertions made by their authors and no admission is made that any reference constitutes prior art. Applicants reserve the right to challenge the accuracy and pertinency of the cited references.

What is Claimed is:

1. A method for haplotyping the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene of an individual, which comprises determining which of the CYP1B1 haplotypes shown in the table immediately below defines one copy of the individual's CYP1B1 gene, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-PS20 on at least one copy of the individual's CYP1B1 gene, and wherein each of the CYP1B1 haplotypes comprises a sequence of polymorphisms whose positions and identities are set forth in the table immediately below:

PS	PS	Haplotype Number(c) (Part 1)									
No.(a)	Position(b)	1	2	3	4	5	6	7	8	9	10
1	1063	C	C	C	C	C	C	C	C	C	C
2	1134	C	T	T	T	T	T	T	T	T	T
3	1342	G	A	G	G	G	G	G	G	G	G
4	1357	T	T	C	C	C	C	C	T	T	T
5	1468	C	C	C	T	T	T	T	C	C	C
6	2454	C	C	C	C	C	C	C	C	C	C
7	2456	C	C	C	C	T	T	T	C	C	C
8	2610	C	C	G	C	G	G	G	C	C	C
9	2823	G	G	G	G	T	T	T	G	G	G
10	3032	C	C	C	C	C	C	C	C	C	C
11	3197	G	G	G	G	G	G	G	G	G	G
12	3551	C	C	C	C	C	C	C	C	C	C
13	6551	T	T	T	T	T	T	T	C	T	T
14	6665	A	A	A	A	A	A	G	A	A	A
15	6798	C	G	G	C	C	C	C	C	C	C
16	6832	C	C	C	C	C	C	C	C	C	C
17	6851	C	T	T	C	C	C	C	C	C	C
18	6862	A	A	A	G	A	A	A	A	A	A
19	7242	G	G	G	G	G	G	G	G	A	G
20	7254	C	C	C	C	C	G	C	C	C	C

(a) PS = polymorphic site;
(b) Position of PS within SEQ ID NO:1;
(c) Alleles for haplotypes are presented 5′ to 3′ in each column;

PS No.(a)	PS Position(b)	Haplotype Number(c) (Part 2) 11	12	13	14	15	16	17	18	19	20
1	1063	C	C	C	C	C	C	C	C	C	T
2	1134	T	T	T	T	T	T	T	T	T	T
3	1342	G	G	G	G	G	G	G	G	G	G
4	1357	T	T	T	T	T	T	T	T	T	T
5	1468	C	C	C	C	C	C	C	C	C	C
6	2454	C	C	C	C	C	C	C	C	C	T
7	2456	C	C	C	C	C	C	T	T	T	C
8	2610	C	C	G	G	G	G	G	G	G	G
9	2823	G	G	G	T	T	T	T	T	T	T
10	3032	C	C	C	A	C	C	C	C	C	C
11	3197	G	G	C	G	C	G	C	G	G	G
12	3551	C	C	C	C	C	C	C	C	T	C
13	6551	T	T	T	T	T	T	T	T	T	T
14	6665	A	A	A	A	A	A	A	A	A	A
15	6798	C	G	G	G	G	G	G	G	C	C
16	6832	C	C	C	C	G	C	C	C	C	C
17	6851	C	T	T	T	T	T	T	T	C	C
18	6862	G	A	A	A	A	A	A	A	A	A
19	7242	G	G	G	G	G	G	G	G	G	G
20	7254	C	C	C	C	C	C	C	C	C	C

(a) PS = polymorphic site;
(b) Position of PS within SEQ ID NO:1;
(c) Alleles for haplotypes are presented 5′ to 3′ in each column.

2. A method for haplotyping the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene of an individual, which comprises determining which of the CYP1B1 haplotype pairs shown in the table immediately below defines both copies of the individual's CYP1B1 gene, wherein the determining step comprises identifying the phased sequence of nucleotides present at each of PS1-PS20 on both copies of the individual's CYP1B1 gene, and wherein each of the CYP1B1 haplotype pairs consists of first and second haplotypes which comprise first and second sequences of polymorphisms whose positions and identities are set forth in the table immediately below:

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 1) 5/5	11/11	10/10	12/12	10/11	11/4	12/18	12/20
1	1063	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	C/C	T/T	T/T	T/T	T/T	T/C	T/T	T/T
5	1468	T/T	C/C	C/C	C/C	C/C	C/T	C/C	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T
7	2456	T/T	C/C	C/C	C/C	C/C	C/C	C/T	C/C
8	2610	G/G	C/C	C/C	C/C	C/C	C/C	C/G	C/G
9	2823	T/T	G/G	G/G	G/G	G/G	G/G	G/T	G/T
10	3032	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	C/C	C/C	C/C	G/G	C/C	C/C	G/G	G/C
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	C/C	C/C	C/C	T/T	C/C	C/C	T/T	T/C
18	6862	A/A	G/G	A/A	A/A	A/G	G/G	A/A	A/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 2) 10/14	10/8	10/3	10/1	11/7	20/13	5/20	11/20
1	1063	C/C	C/C	C/C	C/C	C/C	T/C	C/T	C/T
2	1134	T/T	T/T	T/T	T/C	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/C	T/T	T/C	T/T	C/T	T/T
5	1468	C/C	C/C	C/C	C/C	C/T	C/C	T/C	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	T/C	C/T	C/T
7	2456	C/C	C/C	C/C	C/C	C/T	C/C	T/C	C/C
8	2610	C/G	C/C	C/G	C/C	C/G	G/G	G/G	C/G
9	2823	G/T	G/G	G/G	G/G	G/T	T/G	T/T	G/T
10	3032	C/A	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	G/G	G/G	G/G	G/G	G/G	G/C	G/G	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/C	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/G	A/A	A/A	A/A
15	6798	C/G	C/C	C/G	C/C	C/C	C/G	C/C	C/C
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	C/T	C/C	C/T	C/C	C/C	C/T	C/C	C/C
18	6862	A/A	A/A	A/A	A/A	G/A	A/A	A/A	G/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:1;
(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 3) 12/11	10/17	20/16	17/14	11/13	11/14	12/17	10/12
1	1063	C/C	C/C	T/C	C/C	C/C	C/C	C/C	C/C
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
5	1468	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
6	2454	C/C	C/C	T/C	C/C	C/C	C/C	C/C	C/C
7	2456	C/C	C/T	C/C	T/C	C/C	C/C	C/T	C/C
8	2610	C/C	C/G	G/G	G/G	C/G	C/G	C/G	C/C
9	2823	G/G	G/T	T/T	T/T	G/G	G/T	G/T	G/G
10	3032	C/C	C/C	C/C	C/A	C/C	C/A	C/C	C/C
11	3197	G/G	G/C	G/G	C/G	G/C	G/G	G/C	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	G/C	C/G	C/G	G/G	C/G	C/G	G/G	C/G
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	T/C	C/T	C/T	T/T	C/T	C/T	T/T	C/T
18	6862	A/G	A/A	A/A	A/A	G/A	G/A	A/A	A/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 4) 17/13	10/2	10/9	5/11	12/5	5/6	10/5	19/15
1	1063	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/A	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/T	C/T	T/C	C/C	T/C	T/T
5	1468	C/C	C/C	C/C	T/C	C/T	T/T	C/T	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
7	2456	T/C	C/C	C/C	T/C	C/T	T/T	C/T	T/C
8	2610	G/G	C/C	C/C	G/C	C/G	G/G	C/G	G/G
9	2823	T/G	G/G	G/G	T/G	G/T	T/T	G/T	T/T
10	3032	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	C/C	G/G	G/G	G/G	G/G	G/G	G/G	G/C
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	T/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	G/G	C/G	C/C	C/C	G/C	C/C	C/C	C/G
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/G
17	6851	T/T	C/T	C/C	C/C	T/C	C/C	C/C	C/T
18	6862	A/A	A/A	A/A	A/G	A/A	A/A	A/A	A/A
19	7242	G/G	G/G	G/A	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/G	C/C	C/C

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:1;
(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 5) 12/14
1	1063	C/C
2	1134	T/T
3	1342	G/G
4	1357	T/T
5	1468	C/C
6	2454	C/C
7	2456	C/C
8	2610	C/G
9	2823	G/T
10	3032	C/A
11	3197	G/G
12	3551	C/C
13	6551	T/T
14	6665	A/A
15	6798	G/G
16	6832	C/C
17	6851	T/T
18	6862	A/A
19	7242	G/G
20	7254	C/C

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:1;
(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column.

3. A method for genotyping the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene of an individual, comprising determining for the two copies of the CYP1B1 gene present in the individual the identity of the nucleotide pair at one or more polymorphic sites (PS) selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19, wherein the one or more polymorphic sites (PS) have the position and alternative alleles shown in SEQ ID NO:1.

4. The method of claim 3, wherein the determining step comprises:

(a) isolating from the individual a nucleic acid mixture comprising both copies of the CYP1B1 gene, or a fragment thereof, that are present in the individual;

(b) amplifying from the nucleic acid mixture a target region containing one of the selected polymorphic sites;

(c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region, wherein the oligonucleotide is designed for genotyping the selected polymorphic site in the target region;

(d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized oligonucleotide in the presence of at least one terminator of the reaction, wherein the terminator is complementary to one of the alternative nucleotides present at the selected polymorphic site; and

(e) detecting the presence and identity of the terminator in the extended oligonucleotide.

5. The method of claim 3, which comprises determining for the two copies of the CYP1B1 gene present in the individual the identity of the nucleotide pair at each of PS1-PS20.

6. A method for haplotyping the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene of an individual which comprises determining, for one copy of the CYP1B1 gene present in the individual, the identity of the nucleotide at two or more polymorphic sites (PS) selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19, wherein the selected PS have the position and alternative alleles shown in SEQ ID NO:1.

7. The method of claim 6, further comprising determining the identity of the nucleotide at one or more polymorphic sites selected from the group consisting of PS8, PS9, PS15, PS17, PS18 and PS20, wherein the one or more polymorphic sites (PS) have the position and alternative alleles shown in SEQ ID NO:1.

8. The method of claim 6, wherein the determining step comprises:

(a) isolating from the individual a nucleic acid sample containing only one of the two copies of the CYP1B1 gene, or a fragment thereof, that is present in the individual;

(b) amplifying from the nucleic acid sample a target region containing one of the selected polymorphic sites;

(c) hybridizing a primer extension oligonucleotide to one allele of the amplified target region, wherein the oligonucleotide is designed for haplotyping the selected polymorphic site in the target region;

(d) performing a nucleic acid template-dependent, primer extension reaction on the hybridized oligonucleotide in the presence of at least one terminator of the reaction, wherein the terminator is complementary to one of the alternative nucleotides present at the selected polymorphic site; and

(e) detecting the presence and identity of the terminator in the extended oligonucleotide.

9. A method for predicting a haplotype pair for the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene of an individual comprising:

(a) identifying a CYP1B1 genotype for the individual, wherein the genotype comprises the nucleotide pair at two or more polymorphic sites (PS) selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19, wherein the selected PS have the position and alternative alleles shown in SEQ ID NO:1;

(b) comparing the genotype to the haplotype pair data set forth in the table immediately below; and

(c) determining which haplotype pair is consistent with the genotype of the individual and with the haplotype pair data

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 1) 5/5	11/11	10/10	12/12	10/11	11/4	12/18	12/20
1	1063	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	C/C	T/T	T/T	T/T	T/T	T/C	T/T	T/T
5	1468	T/T	C/C	C/C	C/C	C/C	C/T	C/C	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T
7	2456	T/T	C/C	C/C	C/C	C/C	C/C	C/T	C/C
8	2610	G/G	C/C	C/C	C/C	C/C	C/C	C/G	C/G
9	2823	T/T	G/G	G/G	G/G	G/G	G/G	G/T	G/T
10	3032	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	C/C	C/C	C/C	G/G	C/C	C/C	G/G	G/C
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	C/C	C/C	C/C	T/T	C/C	C/C	T/T	T/C
18	6862	A/A	G/G	A/A	A/A	A/G	G/G	A/A	A/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 2) 10/14	10/8	10/3	10/1	11/7	20/13	5/20	11/20
1	1063	C/C	C/C	C/C	C/C	C/C	T/C	C/T	C/T
2	1134	T/T	T/T	T/T	T/C	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/C	T/T	T/C	T/T	C/T	T/T
5	1468	C/C	C/C	C/C	C/C	C/T	C/C	T/C	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	T/C	C/T	C/T
7	2456	C/C	C/C	C/C	C/C	C/T	C/C	T/C	C/C
8	2610	C/G	C/C	C/G	C/C	C/G	G/G	G/G	C/G
9	2823	G/T	G/G	G/G	G/G	G/T	T/G	T/T	G/T
10	3032	C/A	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	G/G	G/G	G/G	G/G	G/G	G/C	G/G	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/C	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/G	A/A	A/A	A/A
15	6798	C/G	C/C	C/G	C/C	C/C	C/G	C/C	C/C
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	C/T	C/C	C/T	C/C	C/C	C/T	C/C	C/C
18	6862	A/A	A/A	A/A	A/A	G/A	A/A	A/A	G/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:1;
(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 3) 12/11	10/17	20/16	17/14	11/13	11/14	12/17	10/12
1	1063	C/C	C/C	T/C	C/C	C/C	C/C	C/C	C/C
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
5	1468	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
6	2454	C/C	C/C	T/C	C/C	C/C	C/C	C/C	C/C
7	2456	C/C	C/T	C/C	T/C	C/C	C/C	C/T	C/C
8	2610	C/C	C/G	G/G	G/G	C/G	C/G	C/G	C/C
9	2823	G/G	G/T	T/T	T/T	G/G	G/T	G/T	G/G
10	3032	C/C	C/C	C/C	C/A	C/C	C/A	C/C	C/C
11	3197	G/G	G/C	G/G	C/G	G/C	G/G	G/C	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	G/C	C/G	C/G	G/G	C/G	C/G	G/G	C/G
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	T/C	C/T	C/T	T/T	C/T	C/T	T/T	C/T
18	6862	A/G	A/A	A/A	A/A	G/A	G/A	A/A	A/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 4) 17/13	10/2	10/9	5/11	12/5	5/6	10/5	19/15
1	1063	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/A	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/T	C/T	T/C	C/C	T/C	T/T
5	1468	C/C	C/C	C/C	T/C	C/T	T/T	C/T	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
7	2456	T/C	C/C	C/C	T/C	C/T	T/T	C/T	T/C
8	2610	G/G	C/C	C/C	G/C	C/G	G/G	C/G	G/G
9	2823	T/G	G/G	G/G	T/G	G/T	T/T	G/T	T/T
10	3032	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	C/C	G/G	G/G	G/G	G/G	G/G	G/G	G/C
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	T/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	G/G	C/G	C/C	C/C	G/C	C/C	C/C	C/G
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/G
17	6851	T/T	C/T	C/C	C/C	T/C	C/C	C/C	C/T
18	6862	A/A	A/A	A/A	A/G	A/A	A/A	A/A	A/A
19	7242	G/G	G/G	G/A	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/G	C/C	C/C

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:1;
(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 5) 12/14
1	1063	C/C
2	1134	T/T
3	1342	G/G
4	1357	T/T
5	1468	C/C
6	2454	C/C
7	2456	C/C
8	2610	C/G
9	2823	G/T
10	3032	C/A
11	3197	G/G
12	3551	C/C
13	6551	T/T
14	6665	A/A
15	6798	G/G
16	6832	C/C
17	6851	T/T
18	6862	A/A
19	7242	G/G
20	7254	C/C

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:1;
(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column.

10. The method of claim 9, wherein the identified genotype of the individual comprises the nucleotide pair at each of PS1-PS20, which have the position and alternative alleles shown in SEQ ID NO:1.

11. A method for identifying an association between a trait and at least one haplotype or haplotype pair of the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene which comprises comparing the frequency of the haplotype or haplotype pair in a population exhibiting the trait with the frequency of the haplotype or haplotype pair in a reference population, wherein the haplotype is selected from haplotypes 1-20 shown in the table presented immediately below, wherein each of the haplotypes comprises a sequence of polymorphisms whose positions and identities are set forth in the table immediately below:

PS No.(a)	PS Position(b)	Haplotype Number(c) (Part 1) 1	2	3	4	5	6	7	8	9	10
1	1063	C	C	C	C	C	C	C	C	C	C
2	1134	C	T	T	T	T	T	T	T	T	T
3	1342	G	A	G	G	G	G	G	G	G	G
4	1357	T	T	C	C	C	C	C	T	T	T
5	1468	C	C	C	T	T	T	T	C	C	C
6	2454	C	C	C	C	C	C	C	C	C	C
7	2456	C	C	C	C	T	T	T	C	C	C
8	2610	C	C	G	C	G	G	G	C	C	C
9	2823	G	G	G	G	T	T	T	G	G	G
10	3032	C	C	C	C	C	C	C	C	C	C
11	3197	G	G	G	G	G	G	G	G	G	G
12	3551	C	C	C	C	C	C	C	C	C	C
13	6551	T	T	T	T	T	T	T	C	T	T
14	6665	A	A	A	A	A	A	G	A	A	A
15	6798	C	G	G	C	C	C	C	C	C	C
16	6832	C	C	C	C	C	C	C	C	C	C
17	6851	C	T	T	C	C	C	C	C	C	C
18	6862	A	A	A	G	A	A	A	A	A	A
19	7242	G	G	G	G	G	G	G	G	A	G
20	7254	C	C	C	C	C	G	C	C	C	C

PS No.(a)	PS Position(b)	Haplotype Number(c) (Part 2) 11	12	13	14	15	16	17	18	19	20
1	1063	C	C	C	C	C	C	C	C	C	T
2	1134	T	T	T	T	T	T	T	T	T	T
3	1342	G	G	G	G	G	G	G	G	G	G
4	1357	T	T	T	T	T	T	T	T	T	T
5	1468	C	C	C	C	C	C	C	C	C	C
6	2454	C	C	C	C	C	C	C	C	C	T
7	2456	C	C	C	C	C	C	T	T	T	C
8	2610	C	C	G	G	G	G	G	G	G	G
9	2823	G	G	G	T	T	T	T	T	T	T
10	3032	C	C	C	A	C	C	C	C	C	C
11	3197	G	G	C	G	C	G	C	G	G	G
12	3551	C	C	C	C	C	C	C	C	T	C
13	6551	T	T	T	T	T	T	T	T	T	T
14	6665	A	A	A	A	A	A	A	A	A	A
15	6798	C	G	G	G	G	G	G	G	C	C
16	6832	C	C	C	C	G	C	C	C	C	C
17	6851	C	T	T	T	T	T	T	T	C	C
18	6862	G	A	A	A	A	A	A	A	A	A
19	7242	G	G	G	G	G	G	G	G	G	G
20	7254	C	C	C	C	C	C	C	C	C	C

(a) PS = polymorphic site;
(b) Position of PS within SEQ ID NO:1;
(c) Alleles for haplotypes are presented 5′ to 3′ in each column;

and wherein the haplotype pair is selected from the haplotype pairs shown in the table immediately below, wherein each of the CYP1B1 haplotype pairs consists of first and second haplotypes which comprise first and second sequences of polymorphisms whose positions in

SEQ ID NO:1 and identities are set forth in the table immediately below:

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 1) 5/5	11/11	10/10	12/12	10/11	11/4	12/18	12/20
1	1063	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	C/C	T/T	T/T	T/T	T/T	T/C	T/T	T/T
5	1468	T/T	C/C	C/C	C/C	C/C	C/T	C/C	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T
7	2456	T/T	C/C	C/C	C/C	C/C	C/C	C/T	C/C
8	2610	G/G	C/C	C/C	C/C	C/C	C/C	C/G	C/G
9	2823	T/T	G/G	G/G	G/G	G/G	G/G	G/T	G/T
10	3032	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	C/C	C/C	C/C	G/G	C/C	C/C	G/G	G/C
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	C/C	C/C	C/C	T/T	C/C	C/C	T/T	T/C
18	6862	A/A	G/G	A/A	A/A	A/G	G/G	A/A	A/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 2) 10/14	10/8	10/3	10/1	11/7	20/13	5/20	11/20
1	1063	C/C	C/C	C/C	C/C	C/C	T/C	C/T	C/T
2	1134	T/T	T/T	T/T	T/C	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/C	T/T	T/C	T/T	C/T	T/T
5	1468	C/C	C/C	C/C	C/C	C/T	C/C	T/C	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	T/C	C/T	C/T
7	2456	C/C	C/C	C/C	C/C	C/T	C/C	T/C	C/C
8	2610	C/G	C/C	C/G	C/C	C/G	G/G	G/G	C/G
9	2823	G/T	G/G	G/G	G/G	G/T	T/G	T/T	G/T
10	3032	C/A	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	G/G	G/G	G/G	G/G	G/G	G/C	G/G	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/C	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/G	A/A	A/A	A/A
15	6798	C/G	C/C	C/G	C/C	C/C	C/G	C/C	C/C
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	C/T	C/C	C/T	C/C	C/C	C/T	C/C	C/C
18	6862	A/A	A/A	A/A	A/A	G/A	A/A	A/A	G/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:1;
(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 3) 12/11	10/17	20/16	17/14	11/13	11/14	12/17	10/12
1	1063	C/C	C/C	T/C	C/C	C/C	C/C	C/C	C/C
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
5	1468	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
6	2454	C/C	C/C	T/C	C/C	C/C	C/C	C/C	C/C
7	2456	C/C	C/T	C/C	T/C	C/C	C/C	C/T	C/C
8	2610	C/C	C/G	G/G	G/G	C/G	C/G	C/G	C/C
9	2823	G/G	G/T	T/T	T/T	G/G	G/T	G/T	G/G
10	3032	C/C	C/C	C/C	C/A	C/C	C/A	C/C	C/C
11	3197	G/G	G/C	G/G	C/G	G/C	G/G	G/C	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	G/C	C/G	C/G	G/G	C/G	C/G	G/G	C/G
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	T/C	C/T	C/T	T/T	C/T	C/T	T/T	C/T
18	6862	A/G	A/A	A/A	A/A	G/A	G/A	A/A	A/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 4) 17/13	10/2	10/9	5/11	12/5	5/6	10/5	19/15
1	1063	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/A	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/T	C/T	T/C	C/C	T/C	T/T
5	1468	C/C	C/C	C/C	T/C	C/T	T/T	C/T	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
7	2456	T/C	C/C	C/C	T/C	C/T	T/T	C/T	T/C
8	2610	G/G	C/C	C/C	G/C	C/G	G/G	C/G	G/G
9	2823	T/G	G/G	G/G	T/G	G/T	T/T	G/T	T/T
10	3032	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	C/C	G/G	G/G	G/G	G/G	G/G	G/G	G/C
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	T/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	G/G	C/G	C/C	C/C	G/C	C/C	C/C	C/G
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/G
17	6851	T/T	C/T	C/C	C/C	T/C	C/C	C/C	C/T
18	6862	A/A	A/A	A/A	A/G	A/A	A/A	A/A	A/A
19	7242	G/G	G/G	G/A	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/G	C/C	C/C

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:1;
(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 5) 12/14
1	1063	C/C
2	1134	T/T
3	1342	G/G
4	1357	T/T
5	1468	C/C
6	2454	C/C
7	2456	C/C
8	2610	C/G
9	2823	G/T
10	3032	C/A
11	3197	G/G
12	3551	C/C
13	6551	T/T
14	6665	A/A
15	6798	G/G
16	6832	C/C
17	6851	T/T
18	6862	A/A
19	7242	G/G
20	7254	C/C

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:1;
(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;

wherein a higher frequency of the haplotype or haplotype pair in the trait population than in the reference population indicates the trait is associated with the haplotype or haplotype pair.

12. The method of claim 11, wherein the trait is a clinical response to a drug targeting or metabolized by CYP1B1 or to a drug for treating a condition or disease associated with CYP1B1 activity.

13. An isolated oligonucleotide designed for detecting a polymorphism in the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene at a polymorphic site (PS) selected from the group consisting of PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19, wherein the selected PS have the position and alternative alleles shown in SEQ ID NO:1.

14. The isolated oligonucleotide of claim 13, which is an allele-specific oligonucleotide that specifically hybridizes to an allele of the CYP1B1 gene at a region containing the polymorphic site.

15. The allele-specific oligonucleotide of claim 14, which comprises a nucleotide sequence selected from the group consisting of SEQ ID NOS:4-17, the complements of SEQ ID NOS:4-17, and SEQ ID NOS:18-45.

16. The isolated oligonucleotide of claim 13, which is a primer-extension oligonucleotide.

17. The primer-extension oligonucleotide of claim 16,which comprises a nucleotide sequence

selected from the group consisting of SEQ ID NOS:46-73.

18. A kit for haplotyping or genotyping the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) gene of an individual, which comprises a set of oligonucleotides designed to haplotype or genotype each of polymorphic sites (PS) PS1, PS2, PS3, PS4, PS5, PS6, PS7, PS10, PS11, PS12, PS13, PS14, PS16 and PS19, wherein the selected PS have the position and alternative alleles shown in SEQ ID NO:1.

19. The kit of claim 18, which further comprises oligonucleotides designed to genotype or haplotype each of PS8, PS9, PS15, PS17, PS18 and PS20, wherein the selected PS have the position and alternative alleles shown in SEQ ID NO:1.

20. An isolated polynucleotide comprising a nucleotide sequence selected from the group consisting of:

(a) a first nucleotide sequence which comprises a cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) isogene, wherein the CYP1B1 isogene is selected from the group consisting of isogenes 1- 11 and 13 - 20 shown in the table immediately below and wherein each of the isogenes comprises the regions of SEQ ID NO:1 shown in the table immediately below and wherein each of the isogenes 1- 11 and 13 - 20 is further defined by the corresponding sequence of polymorphisms whose positions and identities are set forth in the table immediately below; and

Region	PS	PS	Isogene Number(d) (Part 1)									
Examined(a)	No.(b)	Position(c)	1	2	3	4	5	6	7	8	9	10
882-1955	1	1063	C	C	C	C	C	C	C	C	C	C
882-1955	2	1134	C	T	T	T	T	T	T	T	T	T
882-1955	3	1342	G	A	G	G	G	G	G	G	G	G
882-1955	4	1357	T	T	C	C	C	C	C	T	T	T
882-1955	5	1468	C	C	C	T	T	T	T	C	C	C
2284-3779	6	2454	C	C	C	C	C	C	C	C	C	C
2284-3779	7	2456	C	C	C	C	T	T	T	C	C	C
2284-3779	8	2610	C	C	G	C	G	G	G	C	C	C
2284-3779	9	2823	G	G	G	G	T	T	T	G	G	G
2284-3779	10	3032	C	C	C	C	C	C	C	C	C	C
2284-3779	11	3197	G	G	G	G	G	G	G	G	G	G
2284-3779	12	3551	C	C	C	C	C	C	C	C	C	C
6304-7308	13	6551	T	T	T	T	T	T	T	C	T	T
6304-7308	14	6665	A	A	A	A	A	A	G	A	A	A
6304-7308	15	6798	C	G	G	C	C	C	C	C	C	C
6304-7308	16	6832	C	C	C	C	C	C	C	C	C	C
6304-7308	17	6851	C	T	T	C	C	C	C	C	C	C
6304-7308	18	6862	A	A	A	G	A	A	A	A	A	A
6304-7308	19	7242	G	G	G	G	G	G	G	G	A	G
6304-7308	20	7254	C	C	C	C	C	G	C	C	C	C

Region Examined(a)	PS No.(b)	PS Position(c)	Isogene Number(d) (Part 2) 11	13	14	15	16	17	18	19	20
882-1955	1	1063	C	C	C	C	C	C	C	C	T
882-1955	2	1134	T	T	T	T	T	T	T	T	T
882-1955	3	1342	G	G	G	G	G	G	G	G	G
882-1955	4	1357	T	T	T	T	T	T	T	T	T
882-1955	5	1468	C	C	C	C	C	C	C	C	C
2284-3779	6	2454	C	C	C	C	C	C	C	C	T
2284-3779	7	2456	C	C	C	C	C	T	T	T	C
2284-3779	8	2610	C	G	G	G	G	G	G	G	G
2284-3779	9	2823	G	G	T	T	T	T	T	T	T
2284-3779	10	3032	C	C	A	C	C	C	C	C	C
2284-3779	11	3197	G	C	G	C	G	C	G	G	G
2284-3779	12	3551	C	C	C	C	C	C	C	T	C
6304-7308	13	6551	T	T	T	T	T	T	T	T	T
6304-7308	14	6665	A	A	A	A	A	A	A	A	A
6304-7308	15	6798	C	G	G	G	G	G	G	C	C
6304-7308	16	6832	C	C	C	G	C	C	C	C	C
6304-7308	17	6851	C	T	T	T	T	T	T	C	C
6304-7308	18	6862	G	A	A	A	A	A	A	A	A
6304-7308	19	7242	G	G	G	G	G	G	G	G	G
6304-7308	20	7254	C	C	C	C	C	C	C	C	C

(a) Region examined represents the nucleotide positions defining the start and stop positions within the 1st SEQ ID NO of the sequenced region;
(b) PS = polymorphic site;
(c) Position of PS in SEQ ID NO:1;
(d) Alleles for isogenes are presented 5′ to 3′ in each column;

(b) a second nucleotide sequence which is complementary to the first nucleotide sequence.

21. The isolated polynucleotide of claim 20, which is a DNA molecule and comprises both the first and second nucleotide sequences and further comprises expression regulatory elements operably linked to the first nucleotide sequence.

22. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 21, wherein the organism expresses a CYP1B1 protein that is encoded by the first nucleotide sequence.

23. The recombinant nonhuman organism of claim 22, which is a transgenic animal.

24. An isolated fragment of a cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) isogene, wherein the fragment comprises at least 10 nucleotides in one of the regions of SEQ ID NO:1 shown in the table immediately below and wherein the fragment comprises one or more polymorphisms selected from the group consisting of thymine at PS1, cytosine at PS2, adenine at PS3, cytosine at PS4, thymine at PS5, thymine at PS6, thymine at PS7, adenine at PS10, cytosine at PS11, thymine at PS12, cytosine at PS13, guanine at PS14, guanine at PS16 and adenine at PS19, wherein the selected polymorphism has the position set forth in the table immediately below:

Region Examined(a)	PS No.(b)	PS Position(c)	Isogene Number(d) (Part 1) 1	2	3	4	5	6	7	8	9	10
882-1955	1	1063	C	C	C	C	C	C	C	C	C	C
882-1955	2	1134	C	T	T	T	T	T	T	T	T	T
882-1955	3	1342	G	A	G	G	G	G	G	G	G	G
882-1955	4	1357	T	T	C	C	C	C	C	T	T	T
882-1955	5	1468	C	C	C	T	T	T	T	C	C	C
2284-3779	6	2454	C	C	C	C	C	C	C	C	C	C
2284-3779	7	2456	C	C	C	C	T	T	T	C	C	C
2284-3779	8	2610	C	C	G	C	G	G	G	C	C	C
2284-3779	9	2823	G	G	G	G	T	T	T	G	G	G
2284-3779	10	3032	C	C	C	C	C	C	C	C	C	C
2284-3779	11	3197	G	G	G	G	G	G	G	G	G	G
2284-3779	12	3551	C	C	C	C	C	C	C	C	C	C
6304-7308	13	6551	T	T	T	T	T	T	T	C	T	T
6304-7308	14	6665	A	A	A	A	A	A	G	A	A	A
6304-7308	15	6798	C	G	G	C	C	C	C	C	C	C
6304-7308	16	6832	C	C	C	C	C	C	C	C	C	C
6304-7308	17	6851	C	T	T	C	C	C	C	C	C	C
6304-7308	18	6862	A	A	A	G	A	A	A	A	A	A
6304-7308	19	7242	G	G	G	G	G	G	G	G	A	G
6304-7308	20	7254	C	C	C	C	C	G	C	C	C	C

Region Examined(a)	PS No.(b)	PS Position(c)	Isogene Number(d) (Part 2) 11	13	14	15	16	17	18	19	20
882-1955	1	1063	C	C	C	C	C	C	C	C	T
882-1955	2	1134	T	T	T	T	T	T	T	T	T
882-1955	3	1342	G	G	G	G	G	G	G	G	G
882-1955	4	1357	T	T	T	T	T	T	T	T	T
882-1955	5	1468	C	C	C	C	C	C	C	C	C
2284-3779	6	2454	C	C	C	C	C	C	C	C	T
2284-3779	7	2456	C	C	C	C	C	T	T	T	C
2284-3779	8	2610	C	G	G	G	G	G	G	G	G
2284-3779	9	2823	G	G	T	T	T	T	T	T	T
2284-3779	10	3032	C	C	A	C	C	C	C	C	C
2284-3779	11	3197	G	C	G	C	G	C	G	G	G
2284-3779	12	3551	C	C	C	C	C	C	C	T	C
6304-7308	13	6551	T	T	T	T	T	T	T	T	T
6304-7308	14	6665	A	A	A	A	A	A	A	A	A
6304-7308	15	6798	C	G	G	G	G	G	G	C	C
6304-7308	16	6832	C	C	C	G	C	C	C	C	C
6304-7308	17	6851	C	T	T	T	T	T	T	C	C
6304-7308	18	6862	G	A	A	A	A	A	A	A	A
6304-7308	19	7242	G	G	G	G	G	G	G	G	G
6304-7308	20	7254	C	C	C	C	C	C	C	C	C

(a) Region examined represents the nucleotide positions defining the start and stop positions within SEQ ID NO:1 of the regions sequenced;
(b) PS = polymorphic site;
(c) Position of PS within SEQ ID NO:1;
(d) Alleles for CYP1B1 isogenes are presented 5′ to 3′ in each column.

25. An isolated polynucleotide comprising a coding sequence for a CYP1B1 isogene, wherein the coding sequence comprises the regions of SEQ ID NO:2 that are defined by exons 1-3, except at each of the polymorphic sites which have the positions in SEQ ID NO:2 and polymorphisms set forth in the table immediately below:

PS	PS	Isogene Coding Sequence Number(c) (Part 1)									
No.(a)	Position(b)	1c	3c	4c	5c	6c	7c	8c	9c	10c	11c
8	142	C	G	C	G	G	G	C	C	C	C
9	355	G	G	G	T	T	T	G	G	G	G
10	564	C	C	C	C	C	C	C	C	C	C
11	729	G	G	G	G	G	G	G	G	G	G
13	1047	T	T	T	T	T	T	C	T	T	T
14	1161	A	A	A	A	A	G	A	A	A	A
15	1294	C	G	C	C	C	C	C	C	C	C
16	1328	C	C	C	C	C	C	C	C	C	C
17	1347	C	T	C	C	C	C	C	C	C	C
18	1358	A	A	G	A	A	A	A	A	A	G

PS	PS	Isogene Coding Sequence Number(c) (Part 2)							
No.(a)	Position(b)	13c	14c	15c	16c	17c	18c	19c	20c
8	142	G	G	G	G	G	G	G	G
9	355	G	T	T	T	T	T	T	T
10	564	C	A	C	C	C	C	C	C
11	729	C	G	C	G	C	G	G	G
13	1047	T	T	T	T	T	T	T	T
14	1161	A	A	A	A	A	A	A	A
15	1294	G	G	G	G	G	G	C	C
16	1328	C	C	G	C	C	C	C	C
17	1347	T	T	T	T	T	T	C	C
18	1358	A	A	A	A	A	A	A	A

(a) PS = polymorphic site;
(b) Position of PS in SEQ ID NO:2;
(c) Alleles for the isogene coding sequence are presented 5′ to 3′ in each column; the numerical portion of the isogene coding sequence number represents the number of the parent full CYP1B1 isogene.

26. A recombinant nonhuman organism transformed or transfected with the isolated polynucleotide of claim 25, wherein the organism expresses a cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) protein that is encoded by the polymorphic variant sequence.

27. The recombinant nonhuman organism of claim 26, which is a transgenic animal.

28. An isolated fragment of a CYP1B1 coding sequence, wherein the fragment comprises one or more polymorphisms selected from the group consisting of adenine at a position corresponding to nucleotide 564, cytosine at a position corresponding to nucleotide 729, cytosine at a position corresponding to nucleotide 1047, guanine at a position corresponding to nucleotide 1161 and guanine at a position corresponding to nucleotide 1328 in SEQ ID NO:2.

29 An isolated polypeptide comprising an amino acid sequence which is a polymorphic variant of a

reference sequence for the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) (CYP1B1) protein, wherein the reference sequence comprises SEQ ID NO:3 for the regions encoded by exons 1-3, except the polymorphic variant comprises glycine at a position corresponding to amino acid position 443.

30. An isolated monoclonal antibody specific for and immunoreactive with the isolated polypeptide of claim 29.

31. A method for screening for drugs, or other chemical compounds, that bind to or are enzymatic substrates for the isolated polypeptide of claim 29 which comprises contacting the CYP1B1 polymorphic variant with a candidate agent and assaying for binding or enzymatic activity.

32. An isolated fragment of a CYP1B1 protein, wherein the fragment comprises glycine at a position corresponding to amino acid position 443 in SEQ ID NO:3.

33 A computer system for storing and analyzing polymorphism data for the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile) gene, comprising:

(a) a central processing unit (CPU);

(b) a communication interface;

(c) a display device;

(d) an input device; and

(e) a database containing the polymorphism data;

wherein the polymorphism data comprises any one or more of the haplotypes set forth in the table immediately below:

PS	PS	Haplotype Number(c) (Part 1)									
No.(a)	Position(b)	1	2	3	4	5	6	7	8	9	10
1	1063	C	C	C	C	C	C	C	C	C	C
2	1134	C	T	T	T	T	T	T	T	T	T
3	1342	G	A	G	G	G	G	G	G	G	G
4	1357	T	T	C	C	C	C	C	T	T	T
5	1468	C	C	C	T	T	T	T	C	C	C
6	2454	C	C	C	C	C	C	C	C	C	C
7	2456	C	C	C	C	T	T	T	C	C	C
8	2610	C	C	G	C	G	G	G	C	C	C
9	2823	G	G	G	G	T	T	T	G	G	G
10	3032	C	C	C	C	C	C	C	C	C	C
11	3197	G	G	G	G	G	G	G	G	G	G
12	3551	C	C	C	C	C	C	C	C	C	C
13	6551	T	T	T	T	T	T	T	C	T	T
14	6665	A	A	A	A	A	A	G	A	A	A
15	6798	C	G	G	C	C	C	C	C	C	C
16	6832	C	C	C	C	C	C	C	C	C	C
17	6851	C	T	T	C	C	C	C	C	C	C
18	6862	A	A	A	G	A	A	A	A	A	A
19	7242	G	G	G	G	G	G	G	G	A	G
20	7254	C	C	C	C	C	G	C	C	C	C

PS No.(a)	PS Position(b)	Haplotype Number(c) (Part 2) 11	12	13	14	15	16	17	18	19	20
1	1063	C	C	C	C	C	C	C	C	C	T
2	1134	T	T	T	T	T	T	T	T	T	T
3	1342	G	G	G	G	G	G	G	G	G	G
4	1357	T	T	T	T	T	T	T	T	T	T
5	1468	C	C	C	C	C	C	C	C	C	C
6	2454	C	C	C	C	C	C	C	C	C	T
7	2456	C	C	C	C	C	C	T	T	T	C
8	2610	C	C	G	G	G	G	G	G	G	G
9	2823	G	G	G	T	T	T	T	T	T	T
10	3032	C	C	C	A	C	C	C	C	C	C
11	3197	G	G	C	G	C	G	C	G	G	G
12	3551	C	C	C	C	C	C	C	C	T	C
13	6551	T	T	T	T	T	T	T	T	T	T
14	6665	A	A	A	A	A	A	A	A	A	A
15	6798	C	G	G	G	G	G	G	G	C	C
16	6832	C	C	C	C	G	C	C	C	C	C
17	6851	C	T	T	T	T	T	T	T	C	C
18	6862	G	A	A	A	A	A	A	A	A	A
19	7242	G	G	G	G	G	G	G	G	G	G
20	7254	C	C	C	C	C	C	C	C	C	C

(a) PS = polymorphic site;
(b) Position of PS within SEQ ID NO:1;
(c) Alleles for haplotypes are presented 5′ to 3′ in each column;

the haplotype pairs set forth in the table immediately below:

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 1) 5/5	11/11	10/10	12/12	10/11	11/4	12/18	12/20
1	1063	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	C/C	T/T	T/T	T/T	T/T	T/C	T/T	T/T
5	1468	T/T	C/C	C/C	C/C	C/C	C/T	C/C	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T
7	2456	T/T	C/C	C/C	C/C	C/C	C/C	C/T	C/C
8	2610	G/G	C/C	C/C	C/C	C/C	C/C	C/G	C/G
9	2823	T/T	G/G	G/G	G/G	G/G	G/G	G/T	G/T
10	3032	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	C/C	C/C	C/C	G/G	C/C	C/C	G/G	G/C
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	C/C	C/C	C/C	T/T	C/C	C/C	T/T	T/C
18	6862	A/A	G/G	A/A	A/A	A/G	G/G	A/A	A/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 2) 10/14	10/8	10/3	10/1	11/7	20/13	5/20	11/20
1	1063	C/C	C/C	C/C	C/C	C/C	T/C	C/T	C/T
2	1134	T/T	T/T	T/T	T/C	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/C	T/T	T/C	T/T	C/T	T/T
5	1468	C/C	C/C	C/C	C/C	C/T	C/C	T/C	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	T/C	C/T	C/T
7	2456	C/C	C/C	C/C	C/C	C/T	C/C	T/C	C/C
8	2610	C/G	C/C	C/G	C/C	C/G	G/G	G/G	C/G
9	2823	G/T	G/G	G/G	G/G	G/T	T/G	T/T	G/T
10	3032	C/A	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	G/G	G/G	G/G	G/G	G/G	G/C	G/G	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/C	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/G	A/A	A/A	A/A
15	6798	C/G	C/C	C/G	C/C	C/C	C/G	C/C	C/C
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	C/T	C/C	C/T	C/C	C/C	C/T	C/C	C/C
18	6862	A/A	A/A	A/A	A/A	G/A	A/A	A/A	G/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 3) 12/11	10/17	20/16	17/14	11/13	11/14	12/17	10/12
1	1063	C/C	C/C	T/C	C/C	C/C	C/C	C/C	C/C
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
5	1468	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
6	2454	C/C	C/C	T/C	C/C	C/C	C/C	C/C	C/C
7	2456	C/C	C/T	C/C	T/C	C/C	C/C	C/T	C/C
8	2610	C/C	C/G	G/G	G/G	C/G	C/G	C/G	C/C
9	2823	G/G	G/T	T/T	T/T	G/G	G/T	G/T	G/G
10	3032	C/C	C/C	C/C	C/A	C/C	C/A	C/C	C/C
11	3197	G/G	G/C	G/G	C/G	G/C	G/G	G/C	G/G
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	G/C	C/G	C/G	G/G	C/G	C/G	G/G	C/G
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
17	6851	T/C	C/T	C/T	T/T	C/T	C/T	T/T	C/T
18	6862	A/G	A/A	A/A	A/A	G/A	G/A	A/A	A/A
19	7242	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C

(a) PS = polymorphic site;

(b) Position of PS in SEQ ID NO:1;

(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 4) 17/13	10/2	10/9	5/11	12/5	5/6	10/5	19/15
1	1063	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
2	1134	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
3	1342	G/G	G/A	G/G	G/G	G/G	G/G	G/G	G/G
4	1357	T/T	T/T	T/T	C/T	T/C	C/C	T/C	T/T
5	1468	C/C	C/C	C/C	T/C	C/T	T/T	C/T	C/C
6	2454	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
7	2456	T/C	C/C	C/C	T/C	C/T	T/T	C/T	T/C
8	2610	G/G	C/C	C/C	G/C	C/G	G/G	C/G	G/G
9	2823	T/G	G/G	G/G	T/G	G/T	T/T	G/T	T/T
10	3032	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
11	3197	C/C	G/G	G/G	G/G	G/G	G/G	G/G	G/C
12	3551	C/C	C/C	C/C	C/C	C/C	C/C	C/C	T/C
13	6551	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
14	6665	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A
15	6798	G/G	C/G	C/C	C/C	G/C	C/C	C/C	C/G
16	6832	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/G
17	6851	T/T	C/T	C/C	C/C	T/C	C/C	C/C	C/T
18	6862	A/A	A/A	A/A	A/G	A/A	A/A	A/A	A/A
19	7242	G/G	G/G	G/A	G/G	G/G	G/G	G/G	G/G
20	7254	C/C	C/C	C/C	C/C	C/C	C/G	C/C	C/C

PS No.(a)	PS Position(b)	Haplotype Pair(c) (Part 5) 12/14
1	1063	C/C
2	1134	T/T
3	1342	G/G
4	1357	T/T
5	1468	C/C
6	2454	C/C
7	2456	C/C
8	2610	C/G
9	2823	G/T
10	3032	C/A
11	3197	G/G
12	3551	C/C
13	6551	T/T
14	6665	A/A
15	6798	G/G
16	6832	C/C
17	6851	T/T
18	6862	A/A
19	7242	G/G
20	7254	C/C

(a) PS = polymorphic site;

(b) Position of PS in SEQ ID NO:1;

(c) Haplotype pairs are represented as 1st haplotype/2nd haplotype; with alleles of each haplotype shown 5′ to 3′ as 1st polymorphism/2nd polymorphism in each column;

and the frequency data in Tables 6 and 7.

34. A genome anthology for the cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1

(glaucoma 3, primary infantile) (CYP1B1) gene which comprises two or more CYP1B1 isogenes selected from the group consisting of isogenes 1-20 shown in the table immediately below, and wherein each of the isogenes comprises the regions of SEQ ID NO:1 shown in the table immediately below and wherein each of the isogenes 1-20 is further defined by the corresponding sequence of polymorphisms whose positions and identities are set forth in the table immediately below:

Region	PS	PS	Isogene Number(d) (Part 1)									
Examined(a)	No.(b)	Position(c)	1	2	3	4	5	6	7	8	9	10
882-1955	1	1063	C	C	C	C	C	C	C	C	C	C
882-1955	2	1134	C	T	T	T	T	T	T	T	T	T
882-1955	3	1342	G	A	G	G	G	G	G	G	G	G
882-1955	4	1357	T	T	C	C	C	C	C	T	T	T
882-1955	5	1468	C	C	C	T	T	T	T	C	C	C
2284-3779	6	2454	C	C	C	C	C	C	C	C	C	C
2284-3779	7	2456	C	C	C	C	T	T	T	C	C	C
2284-3779	8	2610	C	C	G	C	G	G	G	C	C	C
2284-3779	9	2823	G	G	G	G	T	T	T	G	G	G
2284-3779	10	3032	C	C	C	C	C	C	C	C	C	C
2284-3779	11	3197	G	G	G	G	G	G	G	G	G	G
2284-3779	12	3551	C	C	C	C	C	C	C	C	C	C
6304-7308	13	6551	T	T	T	T	T	T	T	C	T	T
6304-7308	14	6665	A	A	A	A	A	A	G	A	A	A
6304-7308	15	6798	C	G	G	C	C	C	C	C	C	C
6304-7308	16	6832	C	C	C	C	C	C	C	C	C	C
6304-7308	17	6851	C	T	T	C	C	C	C	C	C	C
6304-7308	18	6862	A	A	A	G	A	A	A	A	A	A
6304-7308	19	7242	G	G	G	G	G	G	G	G	A	G
6304-7308	20	7254	C	C	C	C	C	G	C	C	C	C

(a) Region examined represents the nucleotide positions defining the start and stop positions within SEQ ID NO:1 of the regions sequenced;
(b) PS = polymorphic site;
(c) Position of PS within SEQ ID NO:1;
(d) Alleles for CYP1B1 isogenes are presented 5′ to 3′ in each column;

Region Examined(a)	PS No.(b)	PS Position(c)	Isogene Number(d) (Part 2) 11	12	13	14	15	16	17	18	19	20
882-1955	1	1063	C	C	C	C	C	C	C	C	C	T
882-1955	2	1134	T	T	T	T	T	T	T	T	T	T
882-1955	3	1342	G	G	G	G	G	G	G	G	G	G
882-1955	4	1357	T	T	T	T	T	T	T	T	T	T
882-1955	5	1468	C	C	C	C	C	C	C	C	C	C
2284-3779	6	2454	C	C	C	C	C	C	C	C	C	T
2284-3779	7	2456	C	C	C	C	C	C	T	T	T	C
2284-3779	8	2610	C	C	G	G	G	G	G	G	G	G
2284-3779	9	2823	G	G	G	T	T	T	T	T	T	T
2284-3779	10	3032	C	C	C	A	C	C	C	C	C	C
2284-3779	11	3197	G	G	C	G	C	G	C	G	G	G
2284-3779	12	3551	C	C	C	C	C	C	C	C	T	C
6304-7308	13	6551	T	T	T	T	T	T	T	T	T	T
6304-7308	14	6665	A	A	A	A	A	A	A	A	A	A
6304-7308	15	6798	C	G	G	G	G	G	G	G	C	C
6304-7308	16	6832	C	C	C	C	G	C	C	C	C	C
6304-7308	17	6851	C	T	T	T	T	T	T	T	C	C
6304-7308	18	6862	G	A	A	A	A	A	A	A	A	A
6304-7308	19	7242	G	G	G	G	G	G	G	G	G	G
6304-7308	20	7254	C	C	C	C	C	C	C	C	C	C

(a) Region examined represents the nucleotide positions defining the start and stop positions within SEQ ID NO:1 of the regions sequenced;
(b) PS = polymorphic site;
(c) Position of PS within SEQ ID NO:1;
(d) Alleles for CYP1B1 isogenes are presented 5′ to 3′ in each column.

POLYMORPHISMS IN THE CYP1B1 GENE

```
GGNNNNNNNN NNGNGGTTGG GATACAAGGG TGAAGCCATN GCGCCTAGCC
CCACGTGCAT TTTTTTTTTT AAGGAGGGCA GAAAAGAAGG CATTTGGGCC      100
TCTTATCTGC AATTTTCTTC CGTGTAAACT TATGTGAAGG ATCTGGAGTG
GGACTTGGTG GCTTCAAGCC CTGGCTCCAC TTCTTGATGG CTGCGCGACC      200
TGGGTTAAGT CACGCAACCT CTCTGAACCC TAGTTTATTC ATCTGTAAAC
AGGTCAATAA TAGCACGAGA TTCGCAGCGC GAGTGGAGCT CAAAGTGCAG      300
GGTTGTCCCT GGTGAACCTA TCACTTTTAT ATTTATCCTT TGATGAAGCC
AGTACAATTC CTACCTGGTT AACCAGATAC ATCCCACCTC TTCCCTCGAG      400
TTCGCCCTTC CCCCCGCCTC GTGAAGTCCT TGTTCTCTTA GCTGTCTTGA
AAATCCTATG CATCAGCATG TAGGAAAGGG CGCGCCAGGC GGGGGAAGCC      500
ACCCCCGCCC AAGCGCCTCC GGCTTCCCTT ATAAAGGGAG GGCCCCCTTC
GCGACCGCAA GCGCGCCCAG GAAGACCACA GAGCCGCCGG TGCGCAGCGA      600
GGTGGCGATA CGCGCCCGGG CTCGGCCTGC GGGTGGTGGC CCAAGCGTCC
GCCTCGCTGG CCTGGCAGGC GCGACTGTGC GTGCGCAGCC GAGGGTGGTG      700
GCGGCCGGCA CCCCACGCCA AGGGTGGTGG TGGCCGGCAC CCCACCCTCG
GCCGCCGCCT CCGCGTTTCA GGTGCCGTGA GAAGCGCGGG AGGAGCGGCC      800
GCAGGCAGCG CCCAGGGATA TGACTGGAGC CGACTTTCCA GAAGCGGCGC
ACGCAAAGCC CAGCTCCGCA CGCAAAGGGG AGGCGACAGC AGAAACTTCA      900
ACCCGATAAA GTTCGCCGGG GCGCGGAGAT TCGCCTCCTC CTGCCACTCT
CCGCCCCGCT CGGGTCCCGC CCCGCTAGCT CCCCCAGGCC CCCCCAGTCG     1000
CCCCAGCTTG GCTCCCCGCC CTGCGCCAAC GGCTTCCATC GCAGCCTGGG
CGGCCCCGCG CCCACCAGCG GGCGGCGCCA CCTGGAGTGG CCTCTACGCG     1100
             T
GGAAATCTCA GGGCCAGCTG CGCCCCAGGA GCCTTTGTGT GCCCAAGCAC
                                    C
TGTCGGGGCC CCGGGGCGGG GGAGCGGCTA CTTTTAGGGA TTCCTGATCT     1200
CGCCGCAAGA ACTGGAAAAA ATTTAGCATG CCAAAGAGCC TCCACTGAGG
TGGCAATTTG TTTGCGAGAA CCTAAGATAA AATTTAAACA ACCAACCAGG     1300
GGCGCTGTGA GGCAAACCGC TGCCACTACA CTGGCTTTCC GGGAAGCAAG
                                             A
CTCAAGTCGC GGAGAGGGAA GGGAGGTCGT GCGCTCGGGG CGGGGCGCGC     1400
      C
TCCCAAGTCG AGCGCAGCGG CCGGGGCAGG TTGTACCGAG CGTGGTTCTG
GGGACACCGT GCGGCCTCGA TTGGAGGTGG CTGTGATGAA GCGCGGTTAC     1500
                  T
CGCACAATGG AAACGTGGGC ACCTCCGCTC CCATGAAAGC CTGCTGGTAG
AGCTCCGAGG CCGGCCGGTG CGCCTGGACG GGAGTCCGGG TCAAAGCGGC     1600
CTGGTGTGCG GCGCGCCCCG CCCCCCGCAG GCCCCGCCCT GCCAGGTCGC
GCTGCCCTCC TTCTACCCAG TCCTTAAAAC CCGGAGGAGC GGGATGGCGC     1700
GCTTTGACTC TGGAGTGGGA GTGGGAGCGA GCGCTTCTGC GACTCCAGTT
GTGAGAGCCG CAAGGGCATG GGAATTGACG CCACTCACCG ACCCCCAGTC     1800
TCAATCTCAA CGCTGTGAGG AAACCTCGAC TTTGCCAGGT CCCCAAGGGC
AGCGGGGCTC GGCGAGCGAG GCACCCTTCT CCGTCCCCAT CCCAATCCAA     1900
GCGCTCCTGG CACTGACGAC GCCAAGAGAC TCGAGTGGGA GTTAAAGCTT
CCAGTGAGGG CAGCAGGTGT CCAGGCCGGG CCTGCGGGTT CCTGTTGACG     2000
TCTTGCCCTA GGCAAAGGTC CCAGTTCCTT CTCGGAGCCG GCTGTCCCGC
GCCACTGGAA ACCGCACCTC CCCGCAGGTC AGTCTGTCTG CCGAGGCGCT     2100
GCCCGGCGAC CTCTTCAGAT GGATTATTAC AGGTAGCGGG TGGCGTGGTA
GGTACTTTAA AGGAAATCAA GCGCCACCGC CTCGATGCCC GCAGCGTTGT     2200
CCCCAGATTG CAGGAACCGT TACGCGCCTT GCGGGGAGGG GAAGGGTTTG
```

FIGURE 1A

```
GCGCTGGGTT ACAGCGAGGT GGAAACACGC CCCTTCTCTT CTCCAAGGGA      2300
GAGTGGGTTG GGGATGGGAA GGGGCGTCTT CGGCCATTTC TCCAGAGAGT
CAGCTCCGAC CTCTCCACCC AACGGCACTC AGTCCCCAGA GGCTGGGGTA      2400
GGGGCGTGGG GCGCCCGCTC CTGTCTCTGC ACCCCTGAGT GTCACGCCTT
CTCCTCTCTG TCCCCAGCAT GGGCACCAGC CTCAGCCCGA ACGACCCTTG      2500
   T T
      [exon 2: 2469..
GCCGCTAAAC CCGCTGTCCA TCCAGCAGAC CACGCTCCTG CTACTCCTGT
CGGTGCTGGC CACTGTGCAT GTGGGCCAGC GGCTGCTGAG GCAACGGAGG      2600
CGGCAGCTCC GGTCCGCGCC CCCGGGCCCG TTTGCGTGGC CACTGATCGG
         G
AAACGCGGCG GCGGTGGGCC AGGCGGCTCA CCTCTCGTTC GCTCGCCTGG      2700
CGCGGCGCTA CGGCGACGTT TTCCAGATCC GCCTGGGCAG CTGCCCCATA
GTGGTGCTGA ATGGCGAGCG CGCCATCCAC CAGGCCCTGG TGCAGCAGGG      2800
CTCGGCCTTC GCCGACCGGC CGGCCTTCGC CTCCTTCCGT GTGGTGTCCG
                           T
GCGGCCGCAG CATGGCTTTC GGCCACTACT CGGAGCACTG GAAGGTGCAG      2900
CGGCGCGCAG CCCACAGCAT GATGCGCAAC TTCTTCACGC GCCAGCCGCG
CAGCCGCCAA GTCCTCGAGG GCCACGTGCT GAGCGAGGCG CGCGAGCTGG      3000
TGGCGCTGCT GGTGCGCGGC AGCGCGGACG GCGCCTTCCT CGACCCGAGG
                                    A
CCGCTGACCG TCGTGGCCGT GGCCAACGTC ATGAGTGCCG TGTGTTTCGG      3100
CTGCCGCTAC AGCCACGACG ACCCCGAGTT CCGTGAGCTG CTCAGCCACA
ACGAAGAGTT CGGGCGCACG GTGGGCGCGG GCAGCCTGGT GGACGTGATG      3200
                                                  C
CCCTGGCTGC AGTACTTCCC CAACCCGGTG CGCACCGTTT TCCGCGAATT
CGAGCAGCTC AACCGCAACT TCAGCAACTT CATCCTGGAC AAGTTCTTGA      3300
GGCACTGCGA AAGCCTTCGG CCCGGGGCCG CCCCCCGCGA CATGATGGAC
GCCTTTATCC TCTCTGCGGA AAAGAAGGCG GCCGGGGACT CGCACGGTGG      3400
TGGCGCGCGG CTGGATTTGG AGAACGTACC GGCCACTATC ACTGACATCT
TCGGCGCCAG CCAGGACACC CTGTCCACCG CGCTGCAGTG GCTGCTCCTC      3500
CTCTTCACCA GGTAAAGCCT CTGGGAGGCG TGGGCCAGGT CTTTTCTCCT
                  ..3511]
CTGAAAAAGG CGGAGTAGAG ACAGAATATG CTGAGTTTGC AAGCAGGGCC      3600
T
CCGGGTTTGG GGTTTCGCTC CAGGTCCCCA CCCCTCAAAA CCAAGATCGC
GTCGGTAAAG GGACTCACAG TGAGGGCTGC GACACGCGCA CGCGCCCCAC      3700
CCAGCGGTGC CCCGACCCCT CCGGTCTCCT ATCTTGTCTC TATCGTCCCC
TCCCCTGCTT GCGAGTGAGA ACACATTTGC AAAGACCCCT CCACCCCCCG      3800
GAAAAACAAG AGTTTTTAAA TGCTTGGAGA TGAGCCCTGA TATCTCTCTC
CCTGGCGCAT TACAATCAGA ACTGGAATAG TTCCGAAAGA AAAGGTAATG      3900
TCATAAATAT GTTAAACACA GCAGCCTCTC CTAGGCTAGT CCTCGGCGTG
CATCCGAGGC CGCCCAGCCC TGGCGCTAAA AGCGGGCCGC CCGTCAGGGC      4000
TTTGTTCCAG GCCAAGGAAG CCCATGGAGG CCGGGCCAGC CGACAGGTAA
CCCGCACAGA AACTTTCAGA AGGCGGCCAC AACTAGCGGG CAGCGCTAGG      4100
TTTATAAAAC CTCCGCGCTA GGAGTTTGAG AAATGCCGGG GTAGAAGACA
AGAAGCAGTC ACTTTTACGA AAGCAGAGTA GCATTCAGAA AGGCAGATGG      4200
GATATCCAGG AGGCGCCTGC AGACGTTTCT GGCCCCTGCG CTTGGCTCAG
TTAGCGGACC CCCTGATGCC CACGTTGGTC TCTACTAAGC ACGGATTCAA      4300
CAGGTCCCTG GTGTCGGGTT GCCAGATTTG GCAAAAGAAA AGTAAGTTTT
ACACGGGGAA TACTCACACT AAAAGATTAG CCCTTGTTGA TCTGAAATCC      4400
ATATTTAACT GGGCGGCCTG TAGTATTTAC TGTGGAAACA CTATCCCTAG
```

FIGURE 1B

```
GGGCAAATGT TTCGCAAGGC AAATTTTGAT TGCCGAAGAG ACCAGAAATC      4500
CTGGTTTTGT GTCATTTCTT GGAGCACAAG TGAGCAGTTG GAGATGCTGA
ATCTGCAGGC GCCACAGAAA GGTGTTTGGA AGGCAGAGAA TGACTCTTTC      4600
CTTATTAAAA TCCACTGCAA TCTATATTTC CTTAGATACT GTACAGCTAC
CTTCACAAAT TAAAAGTTTC TGTATACTTA AAATGGCTTT TTAGTATTAA      4700
AATCATAGAA ACACCCATGG TGGGTGAGGG AGAGGCAGAA ATCGAATAAA
GAAAAGTCAA CCAGAGATCA GGGAAAAGGA AATCCCGGAA TAGGTTCCAT      4800
AGGTTTTTGT GCATCCGCAG ATAGGCATTT TAACTTTTGA AACGGCCTTT
GTTTTTCATT AGAACCACAA TAGTTCTCCG AGTACTCCAA TTAGGCGGCA      4900
AAAAGAAATA AAACAATTTG GAGGTCACTT TCAATTCAAT ATGCTGATAC
TTTTTTTTCC TTTTTAGTGA AAAGACGTAT GACAGGGCTT GGCAAATTAC      5000
ACGATCTGTT TTTGTATATA CGTTTTATTG ACGCAGTCAT GCCCATTCAT
TTATACATTG TCGATGGCTG CGTTCACTCT ACAACACAAG GCAGAGCAGA      5100
GTAGTGCAAC AAAGAGGGTT TGGTCCACAA AGTCTAAAAT AGATACTCTC
TGGCTCTCCC CACCATACAT GTTTCTATGA AAAAAGTTGG CAGACTCCTG      5200
CAATATAATG TTGAAGAAGC GATTCTAAAA ATAACTCCAC TTCATCACAT
ACGTGTATAC ATTTTTTTTC GGAGTTGCAA ACAATCAGTT ACTTGTTTTT      5300
TGACTTCTAA CACTTTGACT CAAGGTAGTG GACATTTCTA ATCTTTTAAT
ATTTATTTTG GTAATTTTTG ATGGCTATAT AGTATTTTGC TATGATAATG      5400
TATAGTTATA TATAATCATT TATTGAACTT TAATGTTGGT CATTGGCCTT
GATTGCATTT TTAAAATTTT TATTTTAATT TTATATATTT ACTTACCTTA      5500
GAGACAGGGT CTCACTACCT TACCCAGGCT GGTCCCAAAC TCCTGGGCTC
AAGCAGTTCT CCCGCCGCTG TCTGGCAGGT AGCTGGGGCT ACAGGCGTGT      5600
ACCACCATGC CTGGCTAATT TTTAGAATTT TAGGTCCTGT TGTGGTTACA
TATTTTTGTA TCTCTTAAAT TATTTTCTGA AAATACATTT CTAGGCATGG      5700
ATTACTGGGT TTAAAATGAG GCGGGAAGAG GTTACCTTCA TGTCTCTTGC
CTCGTATTAA GATTTTGTTT TTTAAAAAGA TTGTATCAGT TTGTATCATC      5800
AACAGTGAAT AAGTACTACA GTTTGTACCA TAATTTTATG AACATTGGGT
ATTTGCTTTC AAATTTAAAA AATACAGTAT GTAGCTTTAT CCATCAGGAC      5900
ACCAATTATC TTTGTATAAA ATGAGAACAG CATGTCTGTT GGAATTGTCC
AGGGAAATGA GGGAGAAAAA AAATTTACTT TCACATTGTA ACTTTCGTGG      6000
GCCCTGGGTG CTTTTGCCTT TGTAGATTCC TTATACTATA AAAAAATTAA
AAATTAAATT TCATGACTAC CCTGATATAA AGATGAATGC ATTAAAATGA      6100
TGTATGAAAA TATTTTCTTC TACCTAAAAG TGAGTTTTTT TAGGTCTGAA
GATTGTAAAA GACGTAAAAA TATTTTCATG GGCCCCTAAA AAGTGTTGTG      6200
AGCCCTAGGC ACTGTCCCTG CGGTACCCAA TGGAAAAGTC AGCCTTATCT
ACTATTGTGC TTTTTGAGGC TGGGAAAACT TAAGAGTTTT TTGACTTATA      6300
ATGGGAAAGA CAGCATTAGT CATGCAAGGC CTATTACAGG AAATATAATT
CTTAAAGTCC ATCTTGTAAT TTAGTGAGAA ATTAGGAAGC TGTTTTAGAT      6400
TTTTTTTCCC AGAAATATTA ATTTAGTCAC TGAGCTAGAT AGCCTATTTA
AGAAAAAGTG GAATTAAAAT AAATTATAAT GTGCTTTCTA GATGAAATAA      6500
GAATTTTGCT CACTTGCTTT TCTCTCTCCA CATTAAACAC CAAACAGGTA
     [exon 3: 6548..
TCCTGATGTG CAGACTCGAG TGCAGGCAGA ATTGGATCAG GTCGTGGGGA      6600
C
GGGACCGTCT GCCTTGTATG GGTGACCAGC CCAACCTGCC CTATGTCCTG
GCCTTCCTTT ATGAAGCCAT GCGCTTCTCC AGCTTTGTGC CTGTCACTAT      6700
                G
TCCTCATGCC ACCACTGCCA ACACCTCTGT CTTGGGCTAC CACATTCCCA
AGGACACTGT GGTTTTTGTC AACCAGTGGT CTGTGAATCA TGACCCAGTG      6800
                                                    C
```

FIGURE 1C

```
AAGTGGCCTA ACCCGGAGAA CTTTGATCCA GCTCGATTCT TGGACAAGGA
                                  G
TGGCCTCATC AACAAGGACC TGACCAGCAG AGTGATGATT TTTTCAGTGG      6900
C          G
GCAAAAGGCG GTGCATTGGC GAAGAACTTT CTAAGATGCA GCTTTTTCTC
TTCATCTCCA TCCTGGCTCA CCAGTGCGAT TTCAGGGCCA ACCCAAATGA      7000
GCCTGCGAAA ATGAATTTCA GTTATGGTCT AACCATTAAA CCCAAGTCAT
TTAAAGTCAA TGTCACTCTC AGAGAGTCCA TGGAGCTCCT TGATAGTGCT      7100
GTCCAAAATT TACAAGCCAA GGAAACTTGC CAATAAGAAG CAAGAGGCAA
                ..7136]
GCTGAAATTT TAGAAATATT CACATCTTCG GAGATGAGGA GTAAAATTCA      7200
GTTTTTTTCC AGTTCCTCTT TTGTGCTGCT TCTCAATTAG CGTTTAAGGT
                                           A
GAGCATAAAT CAACTGTCCA TCAGGTGAGG TGTGCTCCAT ACCCAGCGGT      7300
   G
TCTTCATGAG TAGTGGGCTA TGCAGGAGCT TCTGGGAGAT TTTTTTGAGT
CAAAGACTTA AAGGGCCCAA TGAATTATTA TATACATACT GCATCTTGGT      7400   * 7071
TATTTCTGAA GGTAGCATTC TTTGGAGTTA AAATGCACAT ATAGACACAT
ACACCCAAAC ACTTACACCA AACTACTGAA TGAAGAAGTA TTTTGGTAAC      7500
CAGGCCATTT TTGGTGGGAA TCCAAGATTG GTCTCCCATA TGCAGAAATA
GACAAAAAGT ATATTAAACA AAGTTTCAGA GTATATTGTT GAAGAGACAG      7600
AGACAAGTAA TTTCAGTGTA AAGTGTGTGA TTGAAGGTGA TAAGGGAAAA
GATAAAGACC AGAAATTCCC TTTTCACCTT TTCAGGAAAA TAACTTAGAC      7700
TCTAGTATTT ATGGGTGGAT TTATCCTTTT GCCTTCTGGT ATACTTCCTT
ACTTTTAAGG ATAAATCATA AAGTCAGTTG CTCAAAAAGA AATCAATAGT      7800
TGAATTAGTG AGTATAGTGG GGTTCCATGA TTTATCATGA ATTTTAAAGT
ATGCATTATT AAATTGTAAA ACTCCAAGGT GATGTTGTAC CTCTTTTGCT      7900
TGCCAAAGTA CAGAATTTGA ATTATCAGCA AAGAAAAAAA AAAAAGCCAG                * 7073
CCAAGCTTTA AATTATGTGA CCATAATGTA CTGATTTCAG TAAGTCTCAT      8000
AGGTTAAAAA AAAAAGTCAC CAAATAGTGT GAAATATATT ACTTAACTGT                * 7074
CCGTAAGCAG TATATTAGTA TTATCTTGTT CAGGAAAAGG TTGAATAATA      8100
TATGCCTTGT ATAATATTGA AAATTGAAAA GTACAACTAA CGCAACCAAG
TGTGCTAAAA ATGAGCTTGA TTAAATCAAC CACCTATTTT TGACATGGAA      8200
ATGAAGCAGG GTTTCTTTTC TTCACTCAAA TTTTGGCGAA TCTCAAAATT
AGATCCTAAG ATGTGTTCTT ATTTTTATAA CATCTTTATT GAAATTCTAT      8300
TTATAATACA GAATCTTGTT TTGAAAATAA CCTAATTAAT ATATTAAAAT
TCCAAATTCA TGGCATGCTT AAATTTTAAC TAAATTTTAA AGCCATTCTG      8400
ATTATTGAGT TCCAGTTGAA GTTAGTGGAA ATCTGAACAT TCTCCTGTGG
AAGGCAGAGA AATCTAAGCT GTGTCTGCCC AATGAATAAT GGAAAATGCC      8500
ATGAATTACC TGGATGTTCT TTTTACGAGG TGACAAGAGT TGGGGACAGA
ACTCCCATTA CAACTGACCA AGTTTCTCTT CTAGATGATT TTTTGAAAGT      8600
TAACATTAAT GCCTGCTTTT TGGAAAGTCA GAATCAGAAG ATAGTCTTGG
AAGCTGTTTG GAAAAGACAG TGGAGATGAG GTCAGTTGTG TTTTTTAAGA      8700
TGGCAATTAC TTTGGTAGCT GGGAAAGCAT AAAGCTCAAA TGAAATGTAT
GCATTCACAT TTAGAAAAGT GAATTGAAGT TTCAAGTTTT AAAGTTCATT      8800
GCAATTAAAC TTCCAAAGAA AGTTCTACAG TGTCCTAAGT GCTAAGTGCT
TATTACATTT TATTAAGCTT TTTGGAATCT TTGTACCAAA ATTTTAAAAA      8900
AGGGAGTTTT TGATAGTTGT GTGTATGTGT GTGTGGGGTG GGGGGATGGT
AAGAGAAAAG AGAGAAACAC TGAAAAGAAG GAAAGATGGT TAAACATTTT      9000
CCCACTCATT CTGAATTAAT TAATTTGGAG CACAAAATTC AAAGCATGGA
CATTTAGAAG AAAGATGTTT GGCGTAGCAG AGTTAAATCT CAAATAGGCT      9100
ATTAAAAAAG TCTACAACAT AGCAGATCTG TTTTGTGGTT TGGAATATTA
```

FIGURE 1D

```
AAAAACTTCA TGTAATTTTA TTTTAAAATT TCATAGCTGT ACTTCTTGAA      9200
TATAAAAAAT CATGCCAGTA TTTTTAAAGG CATTAGAGTC AACTACACAA
AGCAGGCTTG CCCAGTACAT TTAAATTTTT TGGCACTTGC CATTCCAAAA      9300
TATTATGCCC CACCAAGGCT GAGACAGTGA ATTTGGGCTG CTGTAGCCTA
TTTTTTTAGA TTGAGAAATG TGTAGCTGCA AAAATAATCA TGAACCAATC      9400
TGGATGCCTC ATTATGTCAA CCAGGTCCAG ATGTGCTATA ATCTGTTTTT
ACGTATGTAG GCCCAGTCGT CATCAGATGC TTGCGGCAAA AGGAAAGCTG      9500
TGTTTATATG GAAGAAAGTA AGGTGCTTGG AGTTTACCTG GCTTTTTTAA
TATGCTTATA ACCTAGTTAA AGAAAGGAAA AGAAAACAAA AAACGAATGA      9600
AAATAACTGA ATTTGGAGGC TGGAGTAATC AGATTACTGC TTTAATCAGA
AACCCTCATT GTGTTTCTAC CGGAGAGAGA ATGTATTTGC TGACAACCAT      9700
TAAAGTCAGA AGTTTTACTC CAGGTTATTG CAATAAAGTA TAATGTTTAT
TAAATGCTTC ATTTGTATGT CAAAGCTTTG ACTCTATAAG CAAATTGCTT      9800
TTTTCCAAAA CAAAAAGATG TCTCAGGTTT GTTTTGTGAA TTTTCTAAAA
GCTTTCATGT CCCAGAACTT AGCCTTTACC TGTGAAGTGT TACTACAGCC      9900
TTAATATTTT CCTAGTAGAT CTATATTAGA TCAAATAGTT GCATAGCAGT
ATATGTTAAT TTGTGTGTTT TTAGCTGTGA CACAACTGTG TGATTAAAAG      10000
GTATACTTTA GTAGACATTT ATAACTCAAG GATACCTTCT TATTTAATCT
TTTCTTATTT TTGTACTTTA TCATGAATGC TTTTAGTGTG TGCATAATAG      10100
CTACAGTGCA TAGTTGTAGA CAAAGTACAT TCTGGGGAAA CAACATTTAT
ATGTAGCCTT TACTGTTTGA TATACCAAAT TAAAAAAAAA TTGTATCTCA      10200
TTACTTATAC TGGGACACCA TTACCAAAAT AATAAAAATC ACTTTCATAA
TCTT                                                       10254
```

FIGURE 1E

POLYMORPHISMS IN THE CODING SEQUENCE OF CYP1B1

```
ATGGGCACCA GCCTCAGCCC GAACGACCCT TGGCCGCTAA ACCCGCTGTC
CATCCAGCAG ACCACGCTCC TGCTACTCCT GTCGGTGCTG GCCACTGTGC      100
ATGTGGGCCA GCGGCTGCTG AGGCAACGGA GGCGGCAGCT CCGGTCCGCG
                                                  G
CCCCCGGGCC CGTTTGCGTG GCCACTGATC GGAAACGCGG CGGCGGTGGG      200
CCAGGCGGCT CACCTCTCGT TCGCTCGCCT GGCGCGGCGC TACGGCGACG
TTTTCCAGAT CCGCCTGGGC AGCTGCCCCA TAGTGGTGCT GAATGGCGAG      300
CGCGCCATCC ACCAGGCCCT GGTGCAGCAG GGCTCGGCCT TCGCCGACCG
GCCGGCCTTC GCCTCCTTCC GTGTGGTGTC CGGCGGCCGC AGCATGGCTT      400
    T
TCGGCCACTA CTCGGAGCAC TGGAAGGTGC AGCGGCGCGC AGCCCACAGC
ATGATGCGCA ACTTCTTCAC GCGCCAGCCG CGCAGCCGCC AAGTCCTCGA      500
GGGCCACGTG CTGAGCGAGG CGCGCGAGCT GGTGGCGCTG CTGGTGCGCG
GCAGCGCGGA CGGCGCCTTC CTCGACCCGA GGCCGCTGAC CGTCGTGGCC      600
               A
GTGGCCAACG TCATGAGTGC CGTGTGTTTC GGCTGCCGCT ACAGCCACGA
CGACCCCGAG TTCCGTGAGC TGCTCAGCCA CAACGAAGAG TTCGGGCGCA      700
CGGTGGGCGC GGGCAGCCTG GTGGACGTGA TGCCCTGGCT GCAGTACTTC
                              C
CCCAACCCGG TGCGCACCGT TTTCCGCGAA TTCGAGCAGC TCAACCGCAA      800
CTTCAGCAAC TTCATCCTGG ACAAGTTCTT GAGGCACTGC GAAAGCCTTC
GGCCCGGGGC CGCCCCCCGC GACATGATGG ACGCCTTTAT CCTCTCTGCG      900
GAAAAGAAGG CGGCCGGGGA CTCGCACGGT GGTGGCGCGC GGCTGGATTT
GGAGAACGTA CCGGCCACTA TCACTGACAT CTTCGGCGCC AGCCAGGACA     1000
CCCTGTCCAC CGCGCTGCAG TGGCTGCTCC TCCTCTTCAC CAGGTATCCT
                                                  C
GATGTGCAGA CTCGAGTGCA GGCAGAATTG GATCAGGTCG TGGGGAGGGA     1100
CCGTCTGCCT TGTATGGGTG ACCAGCCCAA CCTGCCCTAT GTCCTGGCCT
TCCTTTATGA AGCCATGCGC TTCTCCAGCT TTGTGCCTGT CACTATTCCT     1200
          G
CATGCCACCA CTGCCAACAC CTCTGTCTTG GGCTACCACA TTCCCAAGGA
CACTGTGGTT TTTGTCAACC AGTGGTCTGT GAATCATGAC CCAGTGAAGT     1300
                                                  C
GGCCTAACCC GGAGAACTTT GATCCAGCTC GATTCTTGGA CAAGGATGGC
                              G                   C
CTCATCAACA AGGACCTGAC CAGCAGAGTG ATGATTTTTT CAGTGGGCAA     1400
      G
AAGGCGGTGC ATTGGCGAAG AACTTTCTAA GATGCAGCTT TTTCTCTTCA
TCTCCATCCT GGCTCACCAG TGCGATTTCA GGGCCAACCC AAATGAGCCT     1500
GCGAAAATGA ATTTCAGTTA TGGTCTAACC ATTAAACCCA AGTCATTTAA
AGTCAATGTC ACTCTCAGAG AGTCCATGGA GCTCCTTGAT AGTGCTGTCC     1600
AAAATTTACA AGCCAAGGAA ACTTGCCAAT AA                        1632
```

FIGURE 2

ISOFORMS OF THE CYP1B1 PROTEIN

```
MGTSLSPNDP WPLNPLSIQQ TTLLLLLSVL ATVHVGQRLL RQRRRQLRSA
                                                  G
PPGPFAWPLI GNAAAVGQAA HLSFARLARR YGDVFQIRLG SCPIVVLNGE      100
RAIHQALVQQ GSAFADRPAF ASFRVVSGGR SMAFGHYSEH WKVQRRAAHS
                    S
MMRNFFTRQP RSRQVLEGHV LSEARELVAL LVRGSADGAF LDPRPLTVVA      200
VANVMSAVCF GCRYSHDDPE FRELLSHNEE FGRTVGAGSL VDVMPWLQYF
PNPVRTVFRE FEQLNRNFSN FILDKFLRHC ESLRPGAAPR DMMDAFILSA      300
EKKAAGDSHG GGARLDLENV PATITDIFGA SQDTLSTALQ WLLLLFTRYP
DVQTRVQAEL DQVVGRDRLP CMGDQPNLPY VLAFLYEAMR FSSFVPVTIP      400
HATTANTSVL GYHIPKDTVV FVNQWSVNHD PVKWPNPENF DPARFLDKDG
                                  L           G
LINKDLTSRV MIFSVGKRRC IGEELSKMQL FLFISILAHQ CDFRANPNEP      500
  S
AKMNFSYGLT IKPKSFKVNV TLRESMELLD SAVQNLQAKE TCQ             543
```

FIGURE 3

CYP1B1.ST25.txt
SEQUENCE LISTING

<110> Genaissance Pharmaceuticals, Inc.
Han, Jin-Hua
Kliem, Stefanie E
Sanchis, Angela

<120> HAPLOTYPES OF THE CYP1B1 GENE

<130> CYP1B1_MWH-1120PCT

<140> TBA
<141> 2001-10-15

<150> 60/240,211
<151> 2000-10-13

<160> 74

<170> PatentIn version 3.1

<210> 1
<211> 10254
<212> DNA
<213> Homo sapiens

<220>
<221> misc_feature
<222> (3)..(12)
<223> n's represent nucleotides of unknown identities

<220>
<221> misc_feature
<222> (14)..(14)
<223> n represents nucleotide of unknown identity

<220>
<221> misc_feature
<222> (40)..(40)
<223> n represents nucleotide of unknown identity

<220>
<221> allele
<222> (1063)..(1063)
<223> PS1: polymorphic base C or T

<220>
<221> allele
<222> (1134)..(1134)
<223> PS2: polymorphic base T or C

<220>
<221> allele
<222> (1342)..(1342)
<223> PS3: polymorphic base G or A

<220>
<221> allele
<222> (1357)..(1357)
<223> PS4: polymorphic base T or C

<220>
<221> allele
<222> (1468)..(1468)
<223> PS5: polymorphic base C or T

<220>
<221> allele
<222> (2454)..(2454)
<223> PS6: polymorphic base C or T

<220>
<221> allele
<222> (2456)..(2456)
<223> PS7: polymorphic base C or T

<220>
<221> allele
<222> (2610)..(2610)
<223> PS8: polymorphic base C or G

<220>
<221> allele
<222> (2823)..(2823)
<223> PS9: polymorphic base G or T

<220>
<221> allele
<222> (3032)..(3032)
<223> PS10: polymorphic base C or A

<220>
<221> allele
<222> (3197)..(3197)
<223> PS11: polymorphic base G or C

<220>
<221> allele
<222> (3551)..(3551)
<223> PS12: polymorphic base C or T

<220>
<221> allele
<222> (6551)..(6551)
<223> PS13: polymorphic base T or C

<220>
<221> allele
<222> (6665)..(6665)
<223> PS14: polymorphic base A or G

<220>
<221> allele
<222> (6798)..(6798)
<223> PS15: polymorphic base G or C

<220>
<221> allele

<222> (6832)..(6832)
<223> PS16: polymorphic base C or G

<220>
<221> allele
<222> (6851)..(6851)
<223> PS17: polymorphic base T or C

<220>
<221> allele
<222> (6862)..(6862)
<223> PS18: polymorphic base A or G

<220>
<221> allele
<222> (7242)..(7242)
<223> PS19: polymorphic base G or A

<220>
<221> allele
<222> (7254)..(7254)
<223> PS20: polymorphic base C or G

<400> 1

ggnnnnnnnn nngnggttgg gatacaaggg tgaagccatn gcgcctagcc ccacgtgcat 60
ttttttttt aaggagggca gaaaagaagg catttgggcc tcttatctgc aattttcttc 120
cgtgtaaact tatgtgaagg atctggagtg ggacttggtg gcttcaagcc ctggctccac 180
ttcttgatgg ctgcgcgacc tgggttaagt cacgcaacct ctctgaaccc tagtttattc 240
atctgtaaac aggtcaataa tagcacgaga ttcgcagcgc gagtggagct caaagtgcag 300
ggttgtccct ggtgaaccta tcacttttat atttatcctt tgatgaagcc agtacaattc 360
ctacctggtt aaccagatac atcccacctc ttccctcgag ttcgcccttc cccccgcctc 420
gtgaagtcct tgttctctta gctgtcttga aaatcctatg catcagcatg taggaaaggg 480
cgcgccaggc gggggaagcc acccccgccc aagcgcctcc ggcttccctt ataaagggag 540
ggcccccttc gcgaccgcaa gcgcgcccag gaagaccaca gagccgccgg tgcgcagcga 600
ggtggcgata cgcgcccggg ctcggcctgc gggtggtggc ccaagcgtcc gcctcgctgg 660
cctggcaggc gcgactgtgc gtgcgcagcc gagggtggtg gcggccggca ccccacgcca 720
agggtggtgg tggccggcac cccaccctcg gccgccgcct ccgcgtttca ggtgccgtga 780
gaagcgcggg aggagcggcc gcaggcagcg cccagggata tgactggagc cgactttcca 840
gaagcggcgc acgcaaagcc cagctccgca cgcaaagggg aggcgacagc agaaacttca 900
acccgataaa gttcgccggg gcgcggagat tcgcctcctc ctgccactct ccgccccgct 960
cgggtcccgc cccgctagct cccccaggcc cccccagtcg ccccagcttg gctccccgcc 1020
ctgcgccaac ggcttccatc gcagcctggg cggccccgcg ccyaccagcg ggcggcgcca 1080
cctggagtgg cctctacgcg ggaaatctca gggccagctg cgccccagga gccyttgtgt 1140
gcccaagcac tgtcggggcc ccggggcggg ggagcggcta cttttaggga ttcctgatct 1200
cgccgcaaga actggaaaaa atttagcatg ccaaagagcc tccactgagg tggcaatttg 1260
tttgcgagaa cctaagataa aatttaaaca accaaccagg ggcgctgtga ggcaaaccgc 1320
tgccactaca ctggctttcc grgaagcaag ctcaagycgc ggagagggaa gggaggtcgt 1380
gcgctcgggg cggggcgcgc tcccaagtcg agcgcagcgg ccggggcagg ttgtaccgag 1440
cgtggttctg gggacaccgt gcggcctyga ttggaggtgg ctgtgatgaa gcgcggttac 1500
cgcacaatgg aaacgtgggc acctccgctc ccatgaaagc ctgctggtag agctccgagg 1560
ccggccggtg cgcctggacg ggagtccggg tcaaagcggc ctggtgtgcg gcgcgccccg 1620
ccccccgcag gccccgccct gccaggtcgc gctgccctcc ttctacccag tccttaaaac 1680
ccggaggagc gggatggcgc gctttgactc tggagtggga gtgggagcga gcgcttctgc 1740
gactccagtt gtgagagccg caagggcatg ggaattgacg ccactcaccg acccccagtc 1800
tcaatctcaa cgctgtgagg aaacctcgac tttgccaggt ccccaagggc agcggggctc 1860
ggcgagcgag gcacccttct ccgtccccat cccaatccaa gcgctcctgg cactgacgac 1920
gccaagagac tcgagtggga gttaaagctt ccagtgaggg cagcaggtgt ccaggccggg 1980
cctgcgggtt cctgttgacg tcttgcccta ggcaaaggtc ccagttcctt ctcggagccg 2040
gctgtcccgc gccactggaa accgcacctc cccgcaggtc agtctgtctg ccgaggcgct 2100
gcccggcgac ctcttcagat ggattattac aggtagcggg tggcgtggta ggtactttaa 2160
aggaaatcaa gcgccaccgc ctcgatgccc gcagcgttgt ccccagattg caggaaccgt 2220
tacgcgcctt gcggggaggg gaagggtttg gcgctgggtt acagcgaggt ggaaacacgc 2280
cccttctctt ctccaaggga gagtgggttg gggatgggaa ggggcgtctt cggccatttc 2340

tccagagagt cagctccgac ctctccaccc aacggcactc agtccccaga ggctggggta 2400
ggggcgtggg gcgcccgctc ctgtctctgc acccctgagt gtcacgcctt ctcytytctg 2460
tccccagcat gggcaccagc ctcagcccga acgacccttg gccgctaaac ccgctgtcca 2520
tccagcagac cacgctcctg ctactcctgt cggtgctggc cactgtgcat gtgggccagc 2580
ggctgctgag gcaacggagg cggcagctcs ggtccgcgcc cccgggcccg tttgcgtggc 2640
cactgatcgg aaacgcggcg gcggtgggcc aggcggctca cctctcgttc gctcgcctgg 2700
cgcggcgcta cggcgacgtt ttccagatcc gcctgggcag ctgccccata gtggtgctga 2760
atggcgagcg cgccatccac caggccctgg tgcagcaggg ctcggccttc gccgaccggc 2820
cgkccttcgc ctccttccgt gtggtgtccg gcggccgcag catggctttc ggccactact 2880
cggagcactg gaaggtgcag cggcgcgcag cccacagcat gatgcgcaac ttcttcacgc 2940
gccagccgcg cagccgccaa gtcctcgagg gccacgtgct gagcgaggcg cgcgagctgg 3000
tggcgctgct ggtgcgcggc agcgcggacg gmgccttcct cgacccgagg ccgctgaccg 3060
tcgtggccgt ggccaacgtc atgagtgccg tgtgtttcgg ctgccgctac agccacgacg 3120
accccgagtt ccgtgagctg ctcagccaca acgaagagtt cgggcgcacg gtgggcgcgg 3180
gcagcctggt ggacgtsatg ccctggctgc agtacttccc caacccggtg cgcaccgttt 3240
tccgcgaatt cgagcagctc aaccgcaact tcagcaactt catcctggac aagttcttga 3300
ggcactgcga aagccttcgg cccggggccg ccccccgcga catgatggac gcctttatcc 3360
tctctgcgga aaagaaggcg gccggggact cgcacggtgg tggcgcgcgg ctggatttgg 3420
agaacgtacc ggccactatc actgacatct tcggcgccag ccaggacacc ctgtccaccg 3480
cgctgcagtg gctgctcctc ctcttcacca ggtaaagcct ctgggaggcg tgggccaggt 3540
cttttctcct ytgaaaaagg cggagtagag acagaatatg ctgagtttgc aagcagggcc 3600
ccgggtttgg ggtttcgctc caggtcccca cccctcaaaa ccaagatcgc gtcggtaaag 3660
ggactcacag tgagggctgc gacacgcgca cgcgccccac ccagcggtgc cccgaccccct 3720
ccggtctcct atcttgtctc tatcgtcccc tcccctgctt gcgagtgaga acacatttgc 3780
aaagacccct ccaccccccg gaaaaacaag agtttttaaa tgcttggaga tgagccctga 3840
tatctctctc cctggcgcat tacaatcaga actggaatag ttccgaaaga aaaggtaatg 3900
tcataaatat gttaaacaca gcagcctctc ctaggctagt cctcggcgtg catccgaggc 3960
cgcccagccc tggcgctaaa agcgggccgc ccgtcagggc tttgttccag gccaaggaag 4020
cccatggagg ccgggccagc cgacaggtaa cccgcacaga aactttcaga aggcggccac 4080
aactagcggg cagcgctagg tttataaaac ctccgcgcta ggagtttgag aaatgccggg 4140
gtagaagaca agaagcagtc acttttacga aagcagagta gcattcagaa aggcagatgg 4200
gatatccagg aggcgcctgc agacgtttct ggcccctgcg cttggctcag ttagcggacc 4260
ccctgatgcc cacgttggtc tctactaagc acggattcaa caggtccctg gtgtcgggtt 4320
gccagatttg gcaaaagaaa agtaagtttt acacggggaa tactcacact aaaagattag 4380
ccctigttga tctgaaatcc atatttaact gggcggcctg tagtatttac tgtggaaaca 4440
ctatccctag gggcaaatgt ttcgcaaggc aaattttgat tgccgaagag accagaaatc 4500
ctggttttgt gtcatttctt ggagcacaag tgagcagttg gagatgctga atctgcaggc 4560
gccacagaaa ggtgtttgga aggcagagaa tgactctttc cttattaaaa tccactgcaa 4620
tctatatttc cttagatact gtacagctac cttcacaaat taaaagtttc tgtatactta 4680
aaatggcttt ttagtattaa aatcatagaa acacccatgg tgggtgaggg agaggcagaa 4740
atcgaataaa gaaaagtcaa ccagagatca gggaaaagga aatcccggaa taggttccat 4800
aggtttttgt gcatccgcag ataggcattt taacttttga aacggccttt gtttttcatt 4860
agaaccacaa tagttctccg agtactccaa ttaggcggca aaaagaaata aaacaatttg 4920
gaggtcactt tcaattcaat atgctgatac tttttttcc tttttagtga aaagacgtat 4980
gacagggctt ggcaaattac acgatctgtt tttgtatata cgttttattg acgcagtcat 5040
gcccattcat ttatacattg tcgatggctg cgttcactct acaacacaag gcagagcaga 5100
gtagtgcaac aaagagggtt tggtccacaa agtctaaaat agatactctc tggctctccc 5160
caccatacat gtttctatga aaaaagttgg cagactcctg caatataatg ttgaagaagc 5220
gattctaaaa ataactccac ttcatcacat acgtgtatac attttttttc ggagttgcaa 5280
acaatcagtt acttgttttt tgacttctaa cactttgact caaggtagtg gacatttcta 5340
atcttttaat atttattttg gtaatttttg atggctatat agtattttgc tatgataatg 5400
tatagttata tataatcatt tattgaactt taatgttggt cattggcctt gattgcattt 5460
ttaaaatttt tattttaatt ttatatattt acttacctta gagacagggt ctcactacct 5520
tacccaggct ggtcccaaac tcctgggctc aagcagttct cccgccgctg tctggcaggt 5580
agctggggct acaggcgtgt accaccatgc ctggctaatt tttagaattt taggtcctgt 5640
tgtggttaca tatttttgta tctcttaaat tattttctga aaatacattt ctaggcatgg 5700
attactgggt ttaaaatgag gcgggaagag gttaccttca tgtctcttgc ctcgtattaa 5760
gattttgttt tttaaaaaga ttgtatcagt ttgtatcatc aacagtgaat aagtactaca 5820
gtttgtacca taatttttatg aacattgggt atttgctttc aaatttaaaa aatacagtat 5880
gtagctttat ccatcaggac accaattatc tttgtataaa atgagaacag catgtctgtt 5940
ggaattgtcc agggaaatga gggagaaaaa aaatttactt tcacattgta actttcgtgg 6000
gccctgggtg cttttgcctt tgtagattcc ttatactata aaaaaattaa aaattaaatt 6060
tcatgactac cctgatataa agatgaatgc attaaaatga tgtatgaaaa tattttcttc 6120
tacctaaaag tgagtttttt taggtctgaa gattgtaaaa gacgtaaaaa tattttcatg 6180
ggcccctaaa aagtgttgtg agccctaggc actgtccctg cggtacccaa tggaaaagtc 6240
agccttatct actattgtgc tttttgaggc tgggaaaact taagagtttt ttgacttata 6300
atgggaaaga cagcattagt catgcaaggc ctattacagg aaatataatt cttaaagtcc 6360
atcttgtaat ttagtgagaa attaggaagc tgttttagat tttttttccc agaaatatta 6420

atttagtcac tgagctagat agcctattta agaaaaagtg gaattaaaat aaattataat 6480
gtgctttcta gatgaaataa gaattttgct cacttgcttt tctctctcca cattaaacac 6540
caaacaggta ycctgatgtg cagactcgag tgcaggcaga attggatcag gtcgtgggga 6600
gggaccgtct gccttgtatg ggtgaccagc ccaacctgcc ctatgtcctg gccttccttt 6660
atgargccat gcgcttctcc agctttgtgc ctgtcactat tcctcatgcc accactgcca 6720
acacctctgt cttgggctac cacattccca aggacactgt ggtttttgtc aaccagtggt 6780
ctgtgaatca tgacccastg aagtggccta acccggagaa ctttgatcca gstcgattct 6840
tggacaagga yggcctcatc arcaaggacc tgaccagcag agtgatgatt ttttcagtgg 6900
gcaaaaggcg gtgcattggc gaagaacttt ctaagatgca gctttttctc ttcatctcca 6960
tcctggctca ccagtgcgat ttcagggcca acccaaatga gcctgcgaaa atgaatttca 7020
gttatggtct aaccattaaa cccaagtcat ttaaagtcaa tgtcactctc agagagtcca 7080
tggagctcct tgatagtgct gtccaaaatt tacaagccaa ggaaacttgc caataagaag 7140
caagaggcaa gctgaaattt tagaaatatt cacatcttcg gagatgagga gtaaaattca 7200
gttttttttcc agttcctctt ttgtgctgct tctcaattag crtttaaggt gagsataaat 7260
caactgtcca tcaggtgagg tgtgctccat acccagcggt tcttcatgag tagtgggcta 7320
tgcaggagct tctgggagat ttttttgagt caaagactta aagggcccaa tgaattatta 7380
tatacatact gcatcttggt tatttctgaa ggtagcattc tttggagtta aaatgcacat 7440
atagacacat acacccaaac acttacacca aactactgaa tgaagaagta ttttggtaac 7500
caggccattt ttggtgggaa tccaagattg gtctcccata tgcagaaata gacaaaaagt 7560
atattaaaca aagtttcaga gtatattgtt gaagagacag agacaagtaa tttcagtgta 7620
aagtgtgtga ttgaaggtga taagggaaaa gataaagacc agaaattccc ttttcacctt 7680
ttcaggaaaa taacttagac tctagtattt atgggtggat ttatcctttt gccttctggt 7740
atacttcctt acttttaagg ataaatcata aagtcagttg ctcaaaaaga aatcaatagt 7800
tgaattagtg agtatagtgg ggttccatga tttatcatga attttaaagt atgcattatt 7860
aaattgtaaa actccaaggt gatgttgtac ctcttttgct tgccaaagta cagaatttga 7920
attatcagca aagaaaaaaa aaaaagccag ccaagcttta aattatgtga ccataatgta 7980
ctgatttcag taagtctcat aggttaaaaa aaaaagtcac caaatagtgt gaaatatatt 8040
acttaactgt ccgtaagcag tatattagta ttatcttgtt caggaaaagg ttgaataata 8100
tatgccttgt ataatattga aaattgaaaa gtacaactaa cgcaaccaag tgtgctaaaa 8160
atgagcttga ttaaatcaac cacctatttt tgacatggaa atgaagcagg gtttcttttc 8220
ttcactcaaa ttttggcgaa tctcaaaatt agatcctaag atgtgttctt atttttataa 8280
catctttatt gaaattctat ttataataca gaatcttgtt ttgaaaataa cctaattaat 8340
atattaaaat tccaaattca tggcatgctt aaatttttaac taaattttaa agccattctg 8400
attattgagt tccagttgaa gttagtggaa atctgaacat tctcctgtgg aaggcagaga 8460
aatctaagct gtgtctgccc aatgaataat ggaaaatgcc atgaattacc tggatgttct 8520
ttttacgagg tgacaagagt tggggacaga actcccatta caactgacca agtttctctt 8580
ctagatgatt ttttgaaagt taacattaat gcctgctttt tggaaagtca gaatcagaag 8640
atagtcttgg aagctgtttg gaaaagacag tggagatgag gtcagttgtg ttttttaaga 8700
tggcaattac tttggtagct gggaaagcat aaagctcaaa tgaaatgtat gcattcacat 8760
ttagaaaagt gaattgaagt ttcaagtttt aaagttcatt gcaattaaac ttccaaagaa 8820
agttctacag tgtcctaagt gctaagtgct tattacattt tattaagctt tttggaatct 8880
ttgtaccaaa attttaaaaa agggagtttt tgatagttgt gtgtatgtgt gtgtggggtg 8940
gggggatggt aagagaaaag agagaaacac tgaaaagaag gaaagatggt taaacatttt 9000
cccactcatt ctgaattaat taatttggag cacaaaattc aaagcatgga catttagaag 9060
aaagatgttt ggcgtagcag agttaaatct caaataggct attaaaaaag tctacaacat 9120
agcagatctg ttttgtggtt tggaatatta aaaaacttca tgtaatttta ttttaaaatt 9180
tcatagctgt acttcttgaa tataaaaaat catgccagta tttttaaagg cattagagtc 9240
aactacacaa agcaggcttg cccagtacat ttaaattttt tggcacttgc cattccaaaa 9300
tattatgccc caccaaggct gagacagtga atttgggctg ctgtagccta ttttttttaga 9360
ttgagaaatg tgtagctgca aaaataatca tgaaccaatc tggatgcctc attatgtcaa 9420
ccaggtccag atgtgctata atctgttttt acgtatgtag gcccagtcgt catcagatgc 9480
ttgcggcaaa aggaaagctg tgtttatatg gaagaaagta aggtgcttgg agtttacctg 9540
gctttttaa tatgcttata acctagttaa agaaaggaaa agaaaacaaa aaacgaatga 9600
aaataactga atttggaggc tggagtaatc agattactgc tttaatcaga aaccctcatt 9660
gtgtttctac cggagagaga atgtatttgc tgacaaccat taaagtcaga agttttactc 9720
caggttattg caataaagta taatgtttat taaatgcttc atttgtatgt caaagctttg 9780
actctataag caaattgctt ttttccaaaa caaaaagatg tctcaggttt gttttgtgaa 9840
ttttctaaaa gctttcatgt cccagaactt agcctttacc tgtgaagtgt tactacagcc 9900
ttaatatttt cctagtagat ctatattaga tcaaatagtt gcatagcagt atatgttaat 9960
ttgtgtgttt ttagctgtga cacaactgtg tgattaaaag gtatacttta gtagacattt 10020
ataactcaag gataccttct tatttaatct tttcttattt ttgtacttta tcatgaatgc 10080
ttttagtgtg tgcataatag ctacagtgca tagttgtaga caaagtacat tctggggaaa 10140
caacatttat atgtagcctt tactgtttga tataccaaat taaaaaaaaa ttgtatctca 10200
ttacttatac tgggacacca ttaccaaaat aataaaaatc actttcataa tctt 10254

<210> 2
<211> 1632

<212> DNA
<213> Homo sapiens

<400> 2
atgggcacca gcctcagccc gaacgaccct tggccgctaa acccgctgtc catccagcag 60
accacgctcc tgctactcct gtcggtgctg gccactgtgc atgtgggcca gcggctgctg 120
aggcaacgga ggcggcagct ccggtccgcg cccccgggcc cgtttgcgtg gccactgatc 180
ggaaacgcgg cggcggtggg ccaggcggct cacctctcgt tcgctcgcct ggcgcggcgc 240
tacggcgacg ttttccagat ccgcctgggc agctgcccca tagtggtgct gaatggcgag 300
cgcgccatcc accaggccct ggtgcagcag ggctcggcct tcgccgaccg gccggccttc 360
gcctccttcc gtgtggtgtc cggcggccgc agcatggctt tcggccacta ctcggagcac 420
tggaaggtgc agcggcgcgc agcccacagc atgatgcgca acttcttcac gcgccagccg 480
cgcagccgcc aagtcctcga gggccacgtg ctgagcgagg cgcgcgagct ggtggcgctg 540
ctggtgcgcg gcagcgcgga cggcgccttc ctcgacccga ggccgctgac cgtcgtggcc 600
gtggccaacg tcatgagtgc cgtgtgtttc ggctgccgct acagccacga cgaccccgag 660
ttccgtgagc tgctcagcca caacgaagag ttcgggcgca cggtgggcgc gggcagcctg 720
gtggacgtga tgccctggct gcagtacttc cccaacccgg tgcgcaccgt tttccgcgaa 780
ttcgagcagc tcaaccgcaa cttcagcaac ttcatcctgg acaagttctt gaggcactgc 840
gaaagccttc ggcccggggc cgccccccgc gacatgatgg acgcctttat cctctctgcg 900
gaaaagaagg cggccgggga ctcgcacggt ggtggcgcgc ggctggattt ggagaacgta 960
ccggccacta tcactgacat cttcggcgcc agccaggaca ccctgtccac cgcgctgcag 1020
tggctgctcc tcctcttcac caggtatcct gatgtgcaga ctcgagtgca ggcagaattg 1080
gatcaggtcg tggggaggga ccgtctgcct tgtatgggtg accagcccaa cctgccctat 1140
gtcctggcct tcctttatga agccatgcgc ttctccagct ttgtgcctgt cactattcct 1200
catgccacca ctgccaacac ctctgtcttg ggctaccaca ttcccaagga cactgtggtt 1260
tttgtcaacc agtggtctgt gaatcatgac ccagtgaagt ggcctaaccc ggagaacttt 1320
gatccagctc gattcttgga caaggatggc ctcatcaaca aggacctgac cagcagagtg 1380
atgatttttt cagtgggcaa aaggcggtgc attggcgaag aactttctaa gatgcagctt 1440
tttctcttca tctccatcct ggctcaccag tgcgatttca gggccaaccc aaatgagcct 1500
gcgaaaatga atttcagtta tggtctaacc attaaaccca agtcatttaa agtcaatgtc 1560
actctcagag agtccatgga gctccttgat agtgctgtcc aaaatttaca agccaaggaa 1620
acttgccaat aa 1632

<210> 3
<211> 543
<212> PRT
<213> Homo sapiens

<400> 3

Met Gly Thr Ser Leu Ser Pro Asn Asp Pro Trp Pro Leu Asn Pro Leu
1 5 10 15

Ser Ile Gln Gln Thr Thr Leu Leu Leu Leu Leu Ser Val Leu Ala Thr
20 25 30

Val His Val Gly Gln Arg Leu Leu Arg Gln Arg Arg Arg Gln Leu Arg
35 40 45

Ser Ala Pro Pro Gly Pro Phe Ala Trp Pro Leu Ile Gly Asn Ala Ala
50 55 60

Ala Val Gly Gln Ala Ala His Leu Ser Phe Ala Arg Leu Ala Arg Arg
65 70 75 80

Tyr Gly Asp Val Phe Gln Ile Arg Leu Gly Ser Cys Pro Ile Val Val
85 90 95

Leu Asn Gly Glu Arg Ala Ile His Gln Ala Leu Val Gln Gln Gly Ser
100 105 110

Ala Phe Ala Asp Arg Pro Ala Phe Ala Ser Phe Arg Val Val Ser Gly
115 120 125

Gly Arg Ser Met Ala Phe Gly His Tyr Ser Glu His Trp Lys Val Gln
130 135 140

Arg Arg Ala Ala His Ser Met Met Arg Asn Phe Phe Thr Arg Gln Pro
145 150 155 160

Arg Ser Arg Gln Val Leu Glu Gly His Val Leu Ser Glu Ala Arg Glu
165 170 175

Leu Val Ala Leu Leu Val Arg Gly Ser Ala Asp Gly Ala Phe Leu Asp
180 185 190

Pro Arg Pro Leu Thr Val Val Ala Val Ala Asn Val Met Ser Ala Val
195 200 205

Cys Phe Gly Cys Arg Tyr Ser His Asp Asp Pro Glu Phe Arg Glu Leu
210 215 220

Leu Ser His Asn Glu Glu Phe Gly Arg Thr Val Gly Ala Gly Ser Leu
225 230 235 240

Val Asp Val Met Pro Trp Leu Gln Tyr Phe Pro Asn Pro Val Arg Thr
245 250 255

Val Phe Arg Glu Phe Glu Gln Leu Asn Arg Asn Phe Ser Asn Phe Ile
260 265 270

Leu Asp Lys Phe Leu Arg His Cys Glu Ser Leu Arg Pro Gly Ala Ala
275 280 285

Pro Arg Asp Met Met Asp Ala Phe Ile Leu Ser Ala Glu Lys Lys Ala
290 295 300

Ala Gly Asp Ser His Gly Gly Gly Ala Arg Leu Asp Leu Glu Asn Val
305 310 315 320

Pro Ala Thr Ile Thr Asp Ile Phe Gly Ala Ser Gln Asp Thr Leu Ser
325 330 335

Thr Ala Leu Gln Trp Leu Leu Leu Leu Phe Thr Arg Tyr Pro Asp Val
340 345 350

Gln Thr Arg Val Gln Ala Glu Leu Asp Gln Val Val Gly Arg Asp Arg
355 360 365

Leu Pro Cys Met Gly Asp Gln Pro Asn Leu Pro Tyr Val Leu Ala Phe
370 375 380

Leu Tyr Glu Ala Met Arg Phe Ser Ser Phe Val Pro Val Thr Ile Pro
385 390 395 400

His Ala Thr Thr Ala Asn Thr Ser Val Leu Gly Tyr His Ile Pro Lys
405 410 415

Asp Thr Val Val Phe Val Asn Gln Trp Ser Val Asn His Asp Pro Val
420 425 430

Lys Trp Pro Asn Pro Glu Asn Phe Asp Pro Ala Arg Phe Leu Asp Lys
435 440 445

Asp Gly Leu Ile Asn Lys Asp Leu Thr Ser Arg Val Met Ile Phe Ser
450 455 460

Val Gly Lys Arg Arg Cys Ile Gly Glu Glu Leu Ser Lys Met Gln Leu
465 470 475 480

Phe Leu Phe Ile Ser Ile Leu Ala His Gln Cys Asp Phe Arg Ala Asn
485 490 495

Pro Asn Glu Pro Ala Lys Met Asn Phe Ser Tyr Gly Leu Thr Ile Lys
500 505 510

Pro Lys Ser Phe Lys Val Asn Val Thr Leu Arg Glu Ser Met Glu Leu
515 520 525

Leu Asp Ser Ala Val Gln Asn Leu Gln Ala Lys Glu Thr Cys Gln
530 535 540

<210> 4
<211> 15
<212> DNA
<213> Homo sapiens

<400> 4
ccgcgccyac cagcg 15

<210> 5
<211> 15
<212> DNA
<213> Homo sapiens

<400> 5
aggagccytt gtgtg 15

<210> 6
<211> 15
<212> DNA
<213> Homo sapiens

<400> 6
ctttccgrga agcaa 15

<210> 7
<211> 15
<212> DNA
<213> Homo sapiens

<400> 7
gctcaagycg cggag 15

<210> 8
<211> 15
<212> DNA
<213> Homo sapiens

<400> 8
gcggcctyga ttgga 15

<210> 9
<211> 15
<212> DNA
<213> Homo sapiens

<400> 9
ccttctcytc tctgt 15

<210> 10
<211> 15
<212> DNA
<213> Homo sapiens

<400> 10

ttctcctytc tgtcc 15

<210> 11
<211> 15
<212> DNA
<213> Homo sapiens

<400> 11
cggacggmgc cttcc 15

<210> 12
<211> 15
<212> DNA
<213> Homo sapiens

<400> 12
tggacgtsat gccct 15

<210> 13
<211> 15
<212> DNA
<213> Homo sapiens

<400> 13
ttctcctytg aaaaa 15

<210> 14
<211> 15
<212> DNA
<213> Homo sapiens

<400> 14
acaggtaycc tgatg 15

<210> 15
<211> 15
<212> DNA
<213> Homo sapiens

<400> 15
tttatgargc catgc 15

<210> 16
<211> 15
<212> DNA
<213> Homo sapiens

<400> 16
gatccagstc gattc 15

<210> 17
<211> 15
<212> DNA
<213> Homo sapiens

<400> 17
aattagcrtt taagg 15

<210> 18
<211> 15

<212> DNA
<213> Homo Sapiens

<400> 18
gcggccccgc gccya 15

<210> 19
<211> 15
<212> DNA
<213> Homo sapiens

<400> 19
gccgcccgct ggtrg 15

<210> 20
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 20
cgccccagga gccyt 15

<210> 21
<211> 15
<212> DNA
<213> Homo sapiens

<400> 21
cttgggcaca caarg 15

<210> 22
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 22
cactggcttt ccgrg 15

<210> 23
<211> 15
<212> DNA
<213> Homo sapiens

<400> 23
ttgagcttgc ttcyc 15

<210> 24
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 24
aagcaagctc aagyc 15

<210> 25
<211> 15
<212> DNA
<213> Homo sapiens

<400> 25
ttccctctcc gcgrc 15

<210> 26
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 26
caccgtgcgg cctyg 15

<210> 27
<211> 15
<212> DNA
<213> Homo sapiens

<400> 27
gccacctcca atcra 15

<210> 28
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 28
gtcacgcctt ctcyt 15

<210> 29
<211> 15
<212> DNA
<213> Homo sapiens

<400> 29
ctggggacag agarg 15

<210> 30
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 30
cacgccttct cctyt 15

<210> 31
<211> 15
<212> DNA
<213> Homo sapiens

<400> 31
tgctggggac agara 15

<210> 32
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 32
gcagcgcgga cggmg 15

<210> 33
<211> 15
<212> DNA

<213> Homo sapiens

<400> 33
ggtcgaggaa ggckc 15

<210> 34
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 34
gcctggtgga cgtsa 15

<210> 35
<211> 15
<212> DNA
<213> Homo sapiens

<400> 35
gcagccaggg catsa 15

<210> 36
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 36
ggtcttttct cctyt 15

<210> 37
<211> 15
<212> DNA
<213> Homo sapiens

<400> 37
tccgcctttt tcara 15

<210> 38
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 38
caccaaacag gtayc 15

<210> 39
<211> 15
<212> DNA
<213> Homo sapiens

<400> 39
tctgcacatc aggrt 15

<210> 40
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 40
ccttccttta tgarg 15

<210> 41
<211> 15
<212> DNA
<213> Homo sapiens

<400> 41
agaagcgcat ggcyt 15

<210> 42
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 42
aactttgatc cagst 15

<210> 43
<211> 15
<212> DNA
<213> Homo sapiens

<400> 43
gtccaagaat cgasc 15

<210> 44
<211> 15
<212> DNA
<213> Homo Sapiens

<400> 44
cttctcaatt agcrt 15

<210> 45
<211> 15
<212> DNA
<213> Homo sapiens

<400> 45
tgctcacctt aaayg 15

<210> 46
<211> 10
<212> DNA
<213> Homo sapiens

<400> 46
gccccgcgcc 10

<210> 47
<211> 10
<212> DNA
<213> Homo sapiens

<400> 47
gcccgctggt 10

<210> 48
<211> 10
<212> DNA
<213> Homo sapiens

<400> 48
cccaggagcc 10

<210> 49
<211> 10
<212> DNA
<213> Homo sapiens

<400> 49
gggcacacaa 10

<210> 50
<211> 10
<212> DNA
<213> Homo sapiens

<400> 50
tggctttccg 10

<210> 51
<211> 10
<212> DNA
<213> Homo sapiens

<400> 51
agcttgcttc 10

<210> 52
<211> 10
<212> DNA
<213> Homo sapiens

<400> 52
caagctcaag 10

<210> 53
<211> 10
<212> DNA
<213> Homo sapiens

<400> 53
cctctccgcg 10

<210> 54
<211> 10
<212> DNA
<213> Homo sapiens

<400> 54
cgtgcggcct 10

<210> 55
<211> 10
<212> DNA
<213> Homo sapiens

<400> 55
acctccaatc 10

<210> 56
<211> 10
<212> DNA
<213> Homo sapiens

<400> 56
acgccttctc 10

<210> 57
<211> 10
<212> DNA
<213> Homo sapiens

<400> 57
gggacagaga 10

<210> 58
<211> 10
<212> DNA
<213> Homo sapiens

<400> 58
gccttctcct 10

<210> 59
<211> 10
<212> DNA
<213> Homo sapiens

<400> 59
tggggacaga 10

<210> 60
<211> 10
<212> DNA
<213> Homo sapiens

<400> 60
gcgcggacgg 10

<210> 61
<211> 10
<212> DNA
<213> Homo sapiens

<400> 61
cgaggaaggc 10

<210> 62
<211> 10
<212> DNA
<213> Homo sapiens

<400> 62
tggtggacgt 10

<210> 63
<211> 10
<212> DNA
<213> Homo sapiens

<400> 63
gccagggcat 10

<210> 64
<211> 10
<212> DNA
<213> Homo sapiens

<400> 64
cttttctcct 10

<210> 65
<211> 10
<212> DNA
<213> Homo sapiens

<400> 65
gccttttca 10

<210> 66
<211> 10
<212> DNA
<213> Homo sapiens

<400> 66
caaacaggta 10

<210> 67
<211> 10
<212> DNA
<213> Homo sapiens

<400> 67
gcacatcagg 10

<210> 68
<211> 10
<212> DNA
<213> Homo sapiens

<400> 68
tcctttatga 10

<210> 69
<211> 10
<212> DNA
<213> Homo sapiens

<400> 69
agcgcatggc 10

<210> 70
<211> 10
<212> DNA
<213> Homo sapiens

<400> 70
tttgatccag 10

<210> 71

<211> 10
<212> DNA
<213> Homo sapiens

<400> 71
caagaatcga 10

<210> 72
<211> 10
<212> DNA
<213> Homo sapiens

<400> 72
ctcaattagc 10

<210> 73
<211> 10
<212> DNA
<213> Homo sapiens

<400> 73
tcaccttaaa 10

<210> 74
<211> 2400
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (30)..(30)
<223> PS1: polymorphic base C or T

<220>
<221> misc_feature
<222> (61)..(120)
<223> n's represent sequence between PS1 and PS2

<220>
<221> allele
<222> (150)..(150)
<223> PS2: polymorphic base T or C

<220>
<221> misc_feature
<222> (181)..(240)
<223> n's represent sequence between PS2 and PS3

<220>
<221> allele
<222> (270)..(270)
<223> PS3: polymorphic base G or A

<220>
<221> misc_feature
<222> (301)..(360)
<223> n's represent sequence between PS3 and PS4

<220>

<221> allele
<222> (390)..(390)
<223> PS4: polymorphic base T or C

<220>
<221> misc_feature
<222> (421)..(480)
<223> n's represent sequence between PS4 and PS5

<220>
<221> allele
<222> (510)..(510)
<223> PS5: polymorphic base C or T

<220>
<221> misc_feature
<222> (541)..(600)
<223> n's represent sequence between PS5 and PS6

<220>
<221> allele
<222> (630)..(630)
<223> PS6: polymorphic base C or T

<220>
<221> misc_feature
<222> (661)..(720)
<223> n's represent sequence between PS6 and PS7

<220>
<221> allele
<222> (750)..(750)
<223> PS7: polymorphic base C or T

<220>
<221> misc_feature
<222> (781)..(840)
<223> n's represent sequence between PS7 and PS8

<220>
<221> allele
<222> (870)..(870)
<223> PS8: polymorphic base C or G

<220>
<221> misc_feature
<222> (901)..(960)
<223> n's represent sequence between PS8 and PS9

<220>
<221> allele
<222> (990)..(990)
<223> PS9: polymorphic base G or T

<220>
<221> misc_feature
<222> (1021)..(1080)

<223> n's represent sequence between PS9 and PS10

<220>
<221> allele
<222> (1110)..(1110)
<223> PS10: polymorphic base C or A

<220>
<221> misc_feature
<222> (1141)..(1200)
<223> n's represent sequence between PS10 and PS11

<220>
<221> allele
<222> (1230)..(1230)
<223> PS11: polymorphic base G or C

<220>
<221> misc_feature
<222> (1261)..(1320)
<223> n's represent sequence between PS11 and PS12

<220>
<221> allele
<222> (1350)..(1350)
<223> PS12: polymorphic base C or T

<220>
<221> misc_feature
<222> (1381)..(1440)
<223> n's represent sequence between PS12 and PS13

<220>
<221> allele
<222> (1470)..(1470)
<223> PS13: polymorphic base T or C

<220>
<221> misc_feature
<222> (1501)..(1560)
<223> n's represent sequence between PS13 and PS14

<220>
<221> allele
<222> (1590)..(1590)
<223> PS14: polymorphic base A or G

<220>
<221> misc_feature
<222> (1621)..(1680)
<223> n's represent sequence between PS14 and PS15

<220>
<221> allele
<222> (1710)..(1710)
<223> PS15: polymorphic base G or C

<220>
<221> misc_feature
<222> (1741)..(1800)
<223> n's represent sequence between PS15 and PS16

<220>
<221> allele
<222> (1830)..(1830)
<223> PS16: polymorphic base C or G

<220>
<221> misc_feature
<222> (1861)..(1920)
<223> n's represent sequence between PS16 and PS17

<220>
<221> allele
<222> (1950)..(1950)
<223> PS17: polymorphic base T or C

<220>
<221> misc_feature
<222> (1981)..(2040)
<223> n's represent sequence between PS17 and PS18

<220>
<221> allele
<222> (2070)..(2070)
<223> PS18: polymorphic base A or G

<220>
<221> misc_feature
<222> (2101)..(2160)
<223> n's represent sequence between PS18 and PS19

<220>
<221> allele
<222> (2190)..(2190)
<223> PS19: polymorphic base G or A

<220>
<221> misc_feature
<222> (2221)..(2280)
<223> n's represent sequence between PS19 and PS20

<220>
<221> allele
<222> (2310)..(2310)
<223> PS20: polymorphic base C or G

<220>
<221> misc_feature
<222> (2341)..(2400)
<223> n's represent sequence 3' to PS20

<400> 74

```
ttccatcgca gcctgggcgg ccccgcgccy accagcgggc ggcgccacct ggagtggcct    60
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn   120
atctcagggc cagctgcgcc ccaggagccy ttgtgtgccc aagcactgtc ggggccccgg   180
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn   240
caaaccgctg ccactacact ggctttccgr gaagcaagct caagtcgcgg agagggaagg   300
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn   360
acactggctt tccgggaagc aagctcaagy cgcggagagg gaagggaggt cgtgcgctcg   420
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn   480
agcgtggttc tggggacacc gtgcggccty gattggaggt ggctgtgatg aagcgcggtt   540
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn   600
ctctgcaccc ctgagtgtca cgccttctcy tctctgtccc cagcatgggc accagcctca   660
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn   720
ctgcaccccc gagtgtcacg ccttctccty tctgtcccca gcatgggcac cagcctcagc   780
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn   840
ggctgctgag gcaacggagg cggcagctcs ggtccgcgcc cccgggcccg tttgcgtggc   900
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn   960
agcagggctc ggccttcgcc gaccggccgk ccttcgcctc cttccgtgtg gtgtccggcg  1020
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  1080
gcgctgctgg tgcgcggcag cgcggacggm gccttcctcg acccgaggcc gctgaccgtc  1140
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  1200
acggtgggcg cgggcagcct ggtggacgts atgccctggc tgcagtactt ccccaacccg  1260
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  1320
tgggaggcgt gggccaggtc ttttctccty tgaaaaaggc ggagtagaga cagaatatgc  1380
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  1440
ctctctccac attaaacacc aaacaggtay cctgatgtgc agactcgagt gcaggcagaa  1500
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  1560
ctgccctatg tcctggcctt cctttatgar gccatgcgct tctccagctt tgtgcctgtc  1620
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  1680
tcaaccagtg gtctgtgaat catgacccas tgaagtggcc taacccggag aactttgatc  1740
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  1800
gtggcctaac ccggagaact ttgatccags tcgattcttg gacaaggatg gcctcatcaa  1860
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  1920
tttgatccag ctcgattctt ggacaaggay ggcctcatca acaaggacct gaccagcaga  1980
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  2040
tcgattcttg gacaaggatg gcctcatcar caaggacctg accagcagag tgatgatttt  2100
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  2160
ttcctctttt gtgctgcttc tcaattagcr tttaaggtga gcataaatca actgtccatc  2220
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  2280
gctgcttctc aattagcgtt taaggtgags ataaatcaac tgtccatcag gtgaggtgtg  2340
nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn  2400
```